POLITECHNIKA GDAŃSKA

Imię i nazwisko autora rozprawy: Krzysztof Kąkol
Dyscyplina naukowa: Informatyka techniczna i telekomunikacja

**ROZPRAWA DOKTORSKA**

Tytuł rozprawy w języku polskim: Poprawa zrozumiałości mowy w obecności zakłóceń z wykorzystaniem efektu Lombarda i profilowania zakłóceń opartego na głębokim uczeniu

Tytuł rozprawy w języku angielskim: Improvement of speech intelligibility in the presence of noise interference using the Lombard effect and an automatic noise interference profiling based on deep learning

| | |
|---|---|
| Promotor<br><br>*podpis* | Drugi promotor<br><br>*podpis* |
| Prof. dr hab. inż. Bożena Kostek | <Tytuł, stopień, imię i nazwisko> |
| Promotor pomocniczy<br><br>*podpis* | Kopromotor<br><br>*podpis* |
| Dr Gražina Korvel | <Tytuł, stopień, imię i nazwisko> |

Gdańsk, rok 2022

GDAŃSK UNIVERSITY OF TECHNOLOGY

The author of the doctoral dissertation: Krzysztof Kąkol
Scientific discipline: Technical Informatics and Telecommunications

**DOCTORAL DISSERTATION**

Title of doctoral dissertation: Improvement of speech intelligibility in the presence of noise interference using the Lombard effect and an automatic noise interference profiling based on deep learning

Title of doctoral dissertation (in Polish): Poprawa zrozumiałości mowy w obecności zakłóceń z wykorzystaniem efektu Lombarda i profilowania zakłóceń opartego na głębokim uczeniu

| | |
|---|---|
| Supervisor<br><br>*signature* | Second supervisor<br><br>*signature* |
| Prof. dr hab. inż. Bożena Kostek | <Title, degree, first name and surname> |
| Auxiliary supervisor<br><br>*signature* | Cosupervisor<br><br>*signature* |
| Dr. Gražina Korvel | <Title, degree, first name and surname> |

Gdańsk, year 2022

POLITECHNIKA GDAŃSKA

## OŚWIADCZENIE

Autor rozprawy doktorskiej: Krzysztof Kąkol

Ja, niżej podpisany(a), oświadczam, iż jestem świadomy(a), że zgodnie z przepisem art. 27 ust. 1 i 2 ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (t.j. Dz.U. z 2021 poz. 1062), uczelnia może korzystać z mojej rozprawy doktorskiej zatytułowanej:

Poprawa zrozumiałości mowy w obecności zakłóceń z wykorzystaniem efektu Lombarda i profilowania zakłóceń opartego na głębokim uczeniu

do prowadzenia badań naukowych lub w celach dydaktycznych.[1]

Świadomy(a) odpowiedzialności karnej z tytułu naruszenia przepisów ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych i konsekwencji dyscyplinarnych określonych w ustawie Prawo o szkolnictwie wyższym i nauce (Dz.U.2021.478 t.j.), a także odpowiedzialności cywilno-prawnej oświadczam, że przedkładana rozprawa doktorska została napisana przeze mnie samodzielnie.

Oświadczam, że treść rozprawy opracowana została na podstawie wyników badań prowadzonych pod kierunkiem i w ścisłej współpracy z promotorem prof. dr hab. inż. Bożeną Kostek, drugim promotorem <drugi promotor>, promotorem pomocniczym dr Gražiną Korvel, kopromotorem <kopromotor>*.

Niniejsza rozprawa doktorska nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadaniem stopnia doktora.

Wszystkie informacje umieszczone w ww. rozprawie uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami, zgodnie z przepisem art. 34 ustawy o prawie autorskim i prawach pokrewnych.

Potwierdzam zgodność niniejszej wersji pracy doktorskiej z załączoną wersją elektroniczną.

Gdańsk, dnia ..... .....
*podpis doktoranta*

Ja, niżej podpisany(a), wyrażam zgodę/~~nie wyrażam zgody~~* na umieszczenie ww. rozprawy doktorskiej w wersji elektronicznej w otwartym, cyfrowym repozytorium instytucjonalnym Politechniki Gdańskiej.

Gdańsk, dnia ..... .....
*podpis doktoranta*

**niepotrzebne usunąć*

[1] Art. 27. 1. Instytucje oświatowe oraz podmioty, o których mowa w art. 7 ust. 1 pkt 1, 2 i 4–8 ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce, mogą na potrzeby zilustrowania treści przekazywanych w celach dydaktycznych lub w celu prowadzenia działalności naukowej korzystać z rozpowszechnionych utworów w oryginale i w tłumaczeniu oraz zwielokrotniać w tym celu rozpowszechnione drobne utwory lub fragmenty większych utworów.
2. W przypadku publicznego udostępniania utworów w taki sposób, aby każdy mógł mieć do nich dostęp w miejscu i czasie przez siebie wybranym korzystanie, o którym mowa w ust. 1, jest dozwolone wyłącznie dla ograniczonego kręgu osób uczących się, nauczających lub prowadzących badania naukowe, zidentyfikowanych przez podmioty wymienione w ust. 1.

POLITECHNIKA GDAŃSKA

**STATEMENT**

The author of the doctoral dissertation: Krzysztof Kąkol

I, the undersigned, declare that I am aware that in accordance with the provisions of Art. 27 (1) and (2) of the Act of 4th February 1994 on Copyright and Related Rights (Journal of Laws of 2021, item 1062), the university may use my doctoral dissertation entitled:
Improvement of speech intelligibility in the presence of noise interference using the Lombard effect and an automatic noise interference profiling based on deep learning
for scientific or didactic purposes.[1]

Gdańsk,..... .....
*signature of the PhD student*

Aware of criminal liability for violations of the Act of 4th February 1994 on Copyright and Related Rights and disciplinary actions set out in the Law on Higher Education and Science (Journal of Laws 2021, item 478), as well as civil liability, I declare, that the submitted doctoral dissertation is my own work.

I declare, that the submitted doctoral dissertation is my own work performed under and in cooperation with the supervision of prof. dr hab.inż. Bożena Kostek, the second supervision of <name of the second supervisor>, the auxiliary supervision of dr Gražina Korvel, the cosupervision of <name of the cosupervisor>*.

This submitted doctoral dissertation has never before been the basis of an official procedure associated with the awarding of a PhD degree.

All the information contained in the above thesis which is derived from written and electronic sources is documented in a list of relevant literature in accordance with Art. 34 of the Copyright and Related Rights Act.

I confirm that this doctoral dissertation is identical to the attached electronic version.

Gdańsk,..... .....
*signature of the PhD student*

I, the undersigned, agree/~~do not agree~~* to include an electronic version of the above doctoral dissertation in the open, institutional, digital repository of Gdańsk University of Technology.

Gdańsk,..... .....
*signature of the PhD student*

**delete where appropriate*

1 Art 27. 1. Educational institutions and entities referred to in art. 7 sec. 1 points 1, 2 and 4–8 of the Act of 20 July 2018 – Law on Higher Education and Science, may use the disseminated works in the original and in translation for the purposes of illustrating the content provided for didactic purposes or in order to conduct research activities, and to reproduce for this purpose disseminated minor works or fragments of larger works.

2. If the works are made available to the public in such a way that everyone can have access to them at the place and time selected by them, as referred to in para. 1, is allowed only for a limited group of people learning, teaching or conducting research, identified by the entities listed in paragraph 1.

POLITECHNIKA GDAŃSKA

# OPIS ROZPRAWY DOKTORSKIEJ

**Autor rozprawy doktorskiej**: Krzysztof Kąkol

**Tytuł rozprawy doktorskiej w języku polskim**: Poprawa zrozumiałości mowy w obecności zakłóceń z wykorzystaniem efektu Lombarda i profilowania zakłóceń opartego na głębokim uczeniu

**Tytuł rozprawy w języku angielskim**: Improvement of speech intelligibility in the presence of noise interference using the Lombard effect and an automatic noise interference profiling based on deep learning

**Język rozprawy doktorskiej**: angielski

**Promotor rozprawy doktorskiej**: prof. dr hab. inż. Bożena Kostek

~~**Drugi promotor rozprawy doktorskiej***: <imię, nazwisko>~~

**Promotor pomocniczy rozprawy doktorskiej***: dr Gražina Korvel

~~**Kopromotor rozprawy doktorskiej***: <imię, nazwisko>~~

**Data obrony**: <dzień, miesiąc, rok>

**Słowa kluczowe rozprawy doktorskiej w języku polskim**: uczenie głębokie, poprawa zrozumiałości mowy, system adaptacyjny, profilowanie zakłóceń

**Słowa kluczowe rozprawy doktorskiej w języku angielskim**: deep learning, speech intelligibility improvement, adaptive system, noise profiling

**Streszczenie rozprawy w języku polskim**:

Efekt Lombarda to zjawisko, dzięki któremu mowa w obecności szumu i zakłóceń może być lepiej zrozumiana. Istnieje wiele charakterystycznych cech takiej mowy, które zostały omówione w niniejszej rozprawie doktorskiej. W pracy zaproponowano stworzenie systemu, który może w czasie rzeczywistym poprawiać jakość i zrozumiałość mowy − mierzone za pomocą metryk obiektywnych i testów subiektywnych. System ten składa się z trzech podstawowych elementów: detekcji typu mowy, profilowania zakłócenia oraz adaptacyjnego systemu doboru najlepszej modyfikacji cech mowy.

Detekcja typu mowy ma na celu wykrycie mowy lombardzkiej w sygnale wejściowym, aby wykluczyć potencjalne modyfikacje sygnału, który posiada cechy mowy lombardzkiej. Drugim elementem jest profilowanie zakłócenia ze względu na istotny wpływ na dobór modyfikacji. Ostatni element systemu to moduł adaptacyjnego doboru modyfikacji sygnału mowy, który na podstawie analizy tego sygnału wybiera rodzaj modyfikacji, która przyniesie największą poprawę jakości mowy w sensie wartości metryki obiektywnej.

W celu rozwiązania postawionych zadań, w niniejszej pracy wykorzystano uczenie maszynowe – w szczególności uczenie głębokie, tj. neuronowe sieci splotowe oraz wielowarstwowe sieci głębokie. W rozprawie doktorskiej wykazano, że jest możliwe stworzenie systemu adaptacyjnego, który będzie poprawiał jakość mowy w obecności zakłócenia.

**Streszczenie rozprawy w języku angielskim**:

The Lombard effect is a phenomenon that results in speech intelligibility improvement when applied to noise. There are many distinctive features of Lombard speech that were recalled in this dissertation. This work proposes the creation of a system capable of improving speech quality and intelligibility in real-time measured by objective metrics and subjective tests. This system consists of three main components: speech type detection, noise profiling, and an adaptive strategy of selection the modification.

The role of the first component is to detect the Lombard speech in the input signal to avoid unnecessary speech modifications when the speech is naturally Lombard in its character. The second module is noise profiling, as the type of noise strongly impacts the selection of the best modification. The last part of the system is the adaptive modification selection component. The selection is made based on the speech signal features, resulting in the most considerable speech quality improvement, measured with objective metrics.

To solve the problem posed, machine learning was used in this dissertation – especially deep learning with convolutional neural networks and typical multilayer networks. It was proven that it is possible to create an adaptive system that would improve speech quality in the presence of noise in real-time or near real-time.

** niepotrzebne skreślić*

POLITECHNIKA GDAŃSKA

# DESCRIPTION OF DOCTORAL DISSERTATION

**The Author of the doctoral dissertation**: Krzysztof Kąkol

**Title of doctoral dissertation**: Improvement of speech intelligibility in the presence of noise interference using the Lombard effect and an automatic noise profiling based on deep learning

**Title of doctoral dissertation in Polish**: Poprawa zrozumiałości mowy w obecności zakłóceń z wykorzystaniem efektu Lombarda i profilowania zakłóceń opartego na głębokim uczeniu

**Language of doctoral dissertation**: English

**Supervisor**: prof. dr hab. inż. Bożena Kostek

~~**Second supervisor***: <first name, surname >~~

**Auxiliary supervisior***: dr Gražina Korvel

~~**Cosupervisior***: <first name, surname >~~

**Date of doctoral defense**: <day, month, year>

**Keywords of doctoral dissertation in Polish**: uczenie głębokie, poprawa zrozumiałości mowy, system adaptacyjny, profilowanie zakłóceń

**Keywords of doctoral dissertation in English**: deep learning, speech intelligibility improvement, adaptive system, noise profiling

**Summary of doctoral dissertation in Polish**:

Efekt Lombarda to zjawisko, dzięki któremu mowa w obecności szumu i zakłóceń może być lepiej zrozumiana. Istnieje wiele charakterystycznych cech takiej mowy, które zostały omówione w niniejszej rozprawie doktorskiej. W pracy zaproponowano stworzenie systemu, który może w czasie rzeczywistym poprawiać jakość i zrozumiałość mowy − mierzone za pomocą metryk obiektywnych i testów subiektywnych. System ten składa się z trzech podstawowych elementów: detekcji typu mowy, profilowania zakłócenia oraz adaptacyjnego systemu doboru najlepszej modyfikacji cech mowy.

Detekcja typu mowy ma na celu wykrycie mowy lombardzkiej w sygnale wejściowym, aby wykluczyć potencjalne modyfikacje sygnału, który posiada cechy mowy lombardzkiej. Drugim elementem jest profilowanie zakłócenia ze względu na istotny wpływ na dobór modyfikacji. Ostatni element systemu to moduł adaptacyjnego doboru modyfikacji sygnału mowy, który na podstawie analizy tego sygnału wybiera rodzaj modyfikacji, która przyniesie największą poprawę jakości mowy w sensie wartości metryki obiektywnej.

W celu rozwiązania postawionych zadań, w niniejszej pracy wykorzystano uczenie maszynowe – w szczególności uczenie głębokie, tj. neuronowe sieci splotowe oraz wielowarstwowe sieci głębokie. W rozprawie doktorskiej wykazano, że jest możliwe stworzenie systemu adaptacyjnego, który będzie poprawiał jakość mowy w obecności zakłócenia.

POLITECHNIKA GDAŃSKA

**Summary of doctoral dissertation in English**:

The Lombard effect is a phenomenon that results in speech intelligibility improvement when applied to noise. There are many distinctive features of Lombard speech that were recalled in this dissertation. This work proposes the creation of a system capable of improving speech quality and intelligibility in real-time measured by objective metrics and subjective tests. This system consists of three main components: speech type detection, noise profiling, and an adaptive strategy of selection the modification.

The role of the first component is to detect the Lombard speech in the input signal to avoid unnecessary speech modifications when the speech is naturally Lombard in its character. The second module is noise profiling, as the type of noise strongly impacts the selection of the best modification. The last part of the system is the adaptive modification selection component. The selection is made based on the speech signal features, resulting in the most considerable speech quality improvement, measured with objective metrics.

To solve the problem posed, machine learning was used in this dissertation – especially deep learning with convolutional neural networks and typical multilayer networks. It was proven that it is possible to create an adaptive system that would improve speech quality in the presence of noise in real-time or near real-time.

**delete where appropriate*

## Streszczenie w j. polskim (rozszerzone)

Efekt Lombarda jest to zjawisko, dzięki któremu mowa w obecności szumu i zakłócenia może być lepiej zrozumiana. Dzieje się to poprzez niezamierzoną modyfikację sposobu mówienia, m.in. poprzez zwiększenie poziomu głosu, co wpływa na większą zrozumiałość mowy w obecności zakłócenia. Istnieje wiele charakterystycznych cech mowy lombardzkiej, takich jak podniesienie częstotliwości podstawowej, wydłużenie samogłosek, podniesienie częstotliwości formantów, przesunięcie energii w wyższe pasma częstotliwości czy spłaszczenie pochylenia widma (Lombard, 1911; Lu and Cooke, 2009; Zollinger and Brumm, 2011).

Efekt Lombarda stał się w dużej mierze inspiracją do przygotowania niniejszej dysertacji. Istnieje wiele potencjalnych zastosowań tego efektu w kontekście przetwarzania sygnału mowy w obecności hałasu, np. w systemach rozgłoszeniowych czy aparatach słuchowych. Zakładając, że efekt Lombarda można wytworzyć za pomocą syntezy mowy, tzn. w taki sposób zmodyfikować sygnał mowy, aby jego charakterystyka odpowiadała mowie lombardzkiej, to można by zastosować tego typu transformacje sygnału w systemach transmisyjnych, aby w relatywnie prosty sposób zwiększyć zrozumiałość mowy. Istotnym elementem takiego systemu musi być działanie adaptacyjne, ponieważ warunki, w których mowa będzie nie tylko dobrze słyszalna, ale też zrozumiała, są różne; różna jest też charakterystyka sygnału mowy.

**W rozprawie doktorskiej zdefiniowano, a następnie udowodniono następujące tezy:**

1. **Zastosowanie splotowych sieci neuronowych (ang. Convolutional Neural Network; CNN) umożliwia wykrycie efektu Lombarda w sygnale mowie z wystarczająco wysoką dokładnością do profilowania rodzaju mowy.**
2. **Wykorzystanie klasycznych metod uczenia maszynowego umożliwia skuteczne i stabilne profilowanie hałasu w czasie zbliżonym do rzeczywistego.**
3. **Istnieje możliwość efektywnego doboru optymalnej metody poprawy zrozumiałości mowy w warunkach zakłócenia, wykorzystując w tym celu profilowanie mowy i hałasu.**

W ramach niniejszej rozprawy doktorskiej przygotowany został system przetwarzania/konwersji wypowiedzi naturalnej w mowę lombardzką w warunkach występowania szumu i zakłóceń. Można zauważyć, że tego typu system powinien być uwarunkowany czasem działania, tj. zaproponowane algorytmy powinny działać w czasie rzeczywistym lub zbliżonym do rzeczywistego. W związku z tym, nie wszystkie cechy mowy lombardzkiej można sztucznie wytworzyć w systemach czasu rzeczywistego. Dlatego w pierwszej kolejności w eksperymentach wytypowano cechy mowy lombardzkiej, które są właściwymi „kandydatami" na ich modyfikację w czasie rzeczywistym – dotyczy to przede wszystkim częstotliwości podstawowej oraz wysokości formantów.

W niniejszej pracy skupiono się na charakterystykach mowy lombardzkiej oraz potencjalnych metodach modyfikacji sygnału mowy, które mogłyby być wykorzystane w syntetyzowanym/sztucznym wytworzeniu efektu Lombarda. Wśród metod omówiono zarówno te, oparte na przetwarzaniu sygnału (PSOLA, model harmoniczny, model sinusoidalny i model źródło-filtr), jak również metodach wykorzystujących uczenie maszynowe, głównie sieci głębokie. Modele głębokie są obecnie wykorzystywane zwłaszcza w systemach przetwarzających tekst na mowę (TTS, Text-to-Speech). Spośród metod opartych na przetwarzaniu sygnału skoncentrowano się głównie na zastosowaniu wokodera WORLD, który – ze względu na relatywną prostotę i niski koszt obliczeniowy – może być wykorzystywany efektywnie w systemach czasu rzeczywistego.

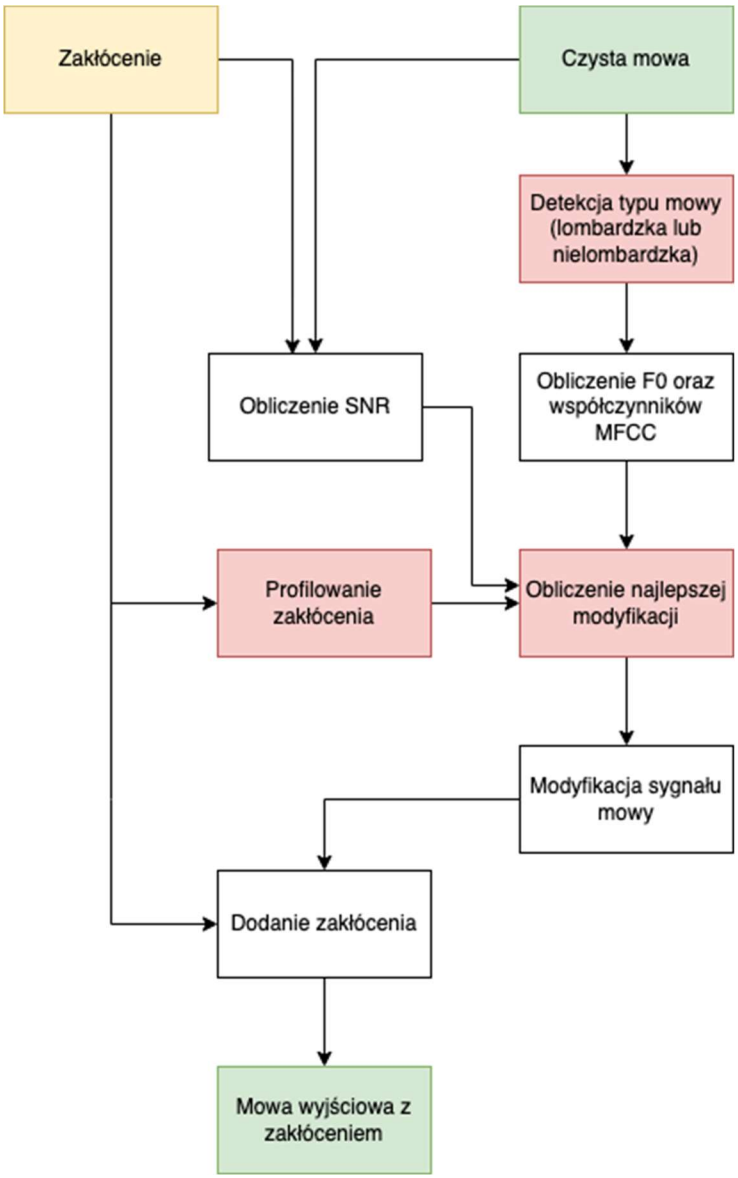
Z punktu widzenia analizy końcowego rozwiązania istotna jest też metoda oceny wyników modyfikacji. Najistotniejszym elementem jest obiektywna i subiektywna ocena zrozumiałości mowy. W systemach teletransmisyjnych stosuje się oba typy pomiarów, dlatego wykorzystano z dwa rodzaje metryk obiektywnych, bazujących na standardach telekomunikacyjnych, publikowanych przez Międzynarodową Unię Telekomunikacyjną (ITU):

- PESQ (ITU-T Recommendation P.862, 2001) - pomiar porównawczy, wymagający zarówno sygnału źródłowego, jak i wyjściowego;

- P.563 (ITU-T Recommendation P.563, 2004) - pomiar jednostronny, wymagający wyłącznie sygnału wyjściowego, a zatem będący znacznie łatwiejszy w zastosowaniach pomiarowych.

Dodatkowo przeprowadzono badania subiektywne, z użyciem metodologii MUSHRA (ITU-T Recommendation BS.1534-1, 2003). Badania te przygotowano w formie ankiety i udostępniono na specjalnej stronie internetowej przy użyciu oprogramowania webMUSHRA (webMUSHRA, 2019).

W toku prac nad niniejszą dysertacją zaprojektowano system, którego celem jest adaptacyjna modyfikacja sygnału mowy w taki sposób, aby w sposób możliwie najbardziej efektywny poprawić jej zrozumiałość. Ponieważ w procesie adaptacji nie jest możliwe wykonanie subiektywnych pomiarów zrozumiałości mowy, bazowano na obiektywnych metrykach (głównie P.563). Schemat niniejszego systemu zaprezentowano na rysunku 1.




*Rysunek 1 Schemat systemu adaptacyjnego poprawiania zrozumiałości mowy w obecności zakłócenia*

W pierwszym kroku system ten dokonuje detekcji rodzaju mowy – tzn. definiuje czy dana próbka jest mową lombardzką czy neutralną. Jeśli zastosowano by modyfikację mowy, która zawiera w sobie cechy mowy lombardzkiej, mogłaby negatywnie wpłynąć na jej jakość, a co za tym idzie, zrozumiałość. Detekcja typu mowy zaimplementowana jest w oparciu o wytrenowaną splotową sieć neuronową. W procesie detekcji przez sieci splotowe wykorzystano reprezentacje 2D sygnału mowy. W ramach niniejszej dysertacji przeprowadzono szereg eksperymentów, które miały na celu porównanie różnych typów wizualizacji 2D, tj. spektrogramów, spektrogramów w skali melowej, chromagramów oraz wizualizacji 2D współczynników mel-cepstralnych (MFCCs, Mel-Frequency Cepstral Coefficients). W toku eksperymentów wykazano, że najlepsze wyniki dają metody oparte na reprezentacji 2D spektrogramów w skali melowej, z dodaniem informacji

o płci mówcy - uzyskano dokładność ok. 98%. W realizacji systemu założono jednak brak wykorzystania informacji o płci, co spowodowało spadek dokładności sieci do poziomu 90%, ale taka dokładność rozpoznawania mowy lombardzkiej jest wystarczająca w rozwiązaniach praktycznych. **Udowodniono w ten sposób tezę nr 1: „Zastosowanie splotowych sieci neuronowych (ang. Convolutional Neural Network; CNN) umożliwia wykrycie efektu Lombarda w sygnale mowie z wystarczająco wysoką dokładnością do profilowania rodzaju mowy".**

Kolejnym istotnym elementem działania systemu jest profilowanie zakłócenia. W toku niniejszej pracy założono, że rodzaj zakłócenia może wpłynąć na typ pożądanej modyfikacji sygnału mowy. Profilowanie zakłócenia wykonano w oparciu o bazę nagrań różnego typu hałasu Aurora (Ellis, 2002). Przyjęto, że model detekcji typu zakłócenia musi być możliwie prosty, a jednocześnie efektywny. Wykonano eksperymenty porównawcze różnych typów modeli uczenia maszynowego i spośród nich − w drodze porównania dokładności modeli wybrano najlepszy – prosty model oparty na twierdzeniu Bayesa (naiwny klasyfikator bayesowski). Jako dane wejściowe do modelu przyjęto uśrednione w krótkim okresie (0,1 sekundy) wartości trzech wybranych charakterystyk sygnału zakłócenia: szerokość pasma widma, płaskość widmową i środek ciężkości widma oraz ich wartości statystycznych. Zaproponowane charakterystyki pozwoliły osiągnąć dokładność profilowania zakłócenia na poziomie 96,76%. Z punktu widzenia działania systemu adaptacyjnej modyfikacji sygnału mowy istotna jest też stabilność detekcji typu zakłócenia, aby nie wprowadzać niepożądanych fluktuacji do komponentu dobierającego optymalną modyfikację. W podsystemie detekcji typu zakłócenia dokonywane jest więc uśrednianie rozpoznawania w dłuższym okresie (np. 1 sekunda), aby zwiększyć odporność algorytmu na błędy rozpoznawania lub krótkotrwałe zmiany cech akustycznych zakłócenia. **Udowodniono tym samym tezę nr 2: „Wykorzystanie klasycznych metod uczenia maszynowego umożliwia skuteczne i stabilne profilowanie hałasu w czasie zbliżonym do rzeczywistego".**

Najważniejszym elementem systemu jest algorytm doboru optymalnego rodzaju modyfikacji sygnału mowy. Wstępnie zaproponowano 24 możliwe modyfikacje sygnału mowy, które obejmują podniesienie wartości częstotliwości podstawowej o 10, 20 lub 30%, a także zwiększenie częstotliwości formantów F1, F2 i F3 o 10, 20 lub 30%. Dla potrzeb eksperymentów częstotliwość podstawową podnoszono przy użyciu programu Praat (Boersma and Weenink, 2018; Corretge, 2012) oraz przy użyciu wokodera WORLD (Morise *et al.*, 2016), zadaptowane przez autora rozprawy. Wartości formantów podnoszono wyłącznie przy użyciu Praat Vocal Toolkit. Docelowo, ze względu na konieczność ograniczenia liczby możliwych wariantów i ze względu na fakt, że zbyt mocne podniesienie częstotliwości formantów wpływa na pogorszenie jakości sygnału mowy (mowa brzmi dużo bardziej nienaturalnie), ograniczono ostatecznie możliwe

modyfikacje formantów do podniesienia ich wartości o 10%. Wytypowano więc 9 możliwych modyfikacji sygnału mowy, zakładających podniesienie F0 o 10, 20 lub 30% z wygładzeniem stosowanym w wokoderze WORLD oraz podniesienie częstotliwości formantów o 10%.

Następnie wykonano liczne eksperymenty, pozwalające wytypować najlepszą możliwą modyfikację dla danego nagrania mowy z wybranego zestawu nagrań, tj. dostępnego w Katedrze Systemów Multimedialnych audio-wizualnego korpusu stworzonego na potrzeby badań nad multimodalnym systemem automatycznego rozpoznawania mowy. Wszystkie nagrania zostały w kolejnym kroku sparametryzowane przez obliczenie częstotliwości podstawowej oraz współczynników MFCC. Na podstawie tych parametrów zbudowano głęboką sieć neuronową, która dokonuje procesu doboru optymalnej modyfikacji. Uzyskana dokładność sieci wynosi – 72%, ale proces klasyfikacji obejmuje dziewięć klas – w tym kontekście należy to uznać za wystarczająco dobry wynik.

Aby potwierdzić powyższą hipotezę, wykonano test dla wszystkich nagrań zawartych w bazie nagrań. Porównano trzy modyfikacje (pierwsze dwie okazały się doświadczalnie najlepsze dla statycznych, nieadaptacyjnych zmian):

- wygładzenie F0 z powiększeniem częstotliwości formantów o 10%,
- podniesienie F0 o 10% z wygładzeniem oraz powiększeniem częstotliwości formantów o 10%,
- adaptacyjna modyfikacja sygnału mowy.

Następnie wykonano eksperymenty, pozwalające porównać wyniki powyższych modyfikacji. Okazało się, że dla męskiego mówcy proces adaptacyjnego doboru modyfikacji jest dużo lepszy niż statyczna modyfikacja. Dla głosu żeńskiego efekt jest porównywalny, ale zaletą takiego systemu adaptacyjnego jest brak konieczności wykonywania detekcji płci mówcy przed wykonaniem doboru modyfikacji. **Udowodniono w ten sposób tezę nr 3: „Istnieje możliwość efektywnego doboru optymalnej metody poprawy zrozumiałości mowy w warunkach zakłócenia, wykorzystując w tym celu profilowanie mowy i hałasu”.**

Z punktu widzenia przetwarzania w czasie rzeczywistym zweryfikowano czas trwania procesów w obrębie wszystkich komponentów systemu. Najdłużej trwającym podprocesem jest detekcja mowy lombardzkiej, ale proces detekcji trwa ok. 2 razy krócej niż próbka. Jest to więc czas, który pozwala na realizację procesu doboru optymalnej modyfikacji w czasie rzeczywistym. Pozostałe podprocesy zajmują od kilku do kilkudziesięciu milisekund. Należy zaznaczyć, że pomiary były wykonywane na zwykłym komputerze przenośnym, bez wspomagania układów GPU.

W literaturze istnieją opisy metod i implementacji modyfikacji sygnału mowy w taki sposób, by wykorzystać efekt mowy lombardzkiej (Bollepalli et al., 2019). Metody te zwykle

wykorzystywane są w systemach text-to-speech, a zatem konieczna jest znajomość zawartości semantycznej lub fonemowej tekstu. Niniejsza dysertacja pokrywa inny obszar zastosowań poprawy jakości mowy – w sytuacji, gdy fonemowa, semantyczna ani syntaktyczna zawartość tekstu nie jest znana. W takim przypadku mowa jest wyłącznie sygnałem dźwiękowym, dla którego jakość i zrozumiałość powinny być poprawiona. Niniejsza praca może być więc istotnym wkładem w badania nad możliwościami poprawy zrozumiałości mowy w obecności hałasu.

**W niniejszej pracy można wyróżnić następujące oryginalne osiągnięcia autora:**

1. Zaproponowano i zaimplementowano nowatorską metodę detekcji mowy lombardzkiej przy użyciu splotowych sieci neuronowych (ang. CNN).
2. Wykonano szereg eksperymentów dotyczących wykorzystania różnych reprezentacji 2D sygnału mowy w kontekście detekcji mowy lombardzkiej.
3. Zaproponowano i zaimplementowano efektywną i nowatorską metodę profilowania zakłócenia przy użyciu metod uczenia maszynowego.
4. Wykonano szereg eksperymentów dotyczących poprawiania jakości mowy oraz zbadano możliwości poprawy jakości mowy przy użyciu wartości częstotliwości podstawowej oraz formantów. Obliczono zarówno wartości metryk obiektywnych, jak i zbadano wpływ zmian na zrozumiałość mowy przy użyciu testów subiektywnych.
5. Zaproponowano i zaimplementowano nowatorską metodę adaptacyjnego doboru najlepszej możliwej modyfikacji sygnału mowy w obecności zakłócenia w oparciu o cechy samego sygnału mowy oraz profilu zakłócenia.

**Dalsze kierunki badań**

System może być wykorzystany efektywnie w wielu zastosowaniach, np. w aparatach słuchowych, w których istnieje praktyczna potrzeba zwiększenia zrozumiałości mowy w obecności szumu i zakłóceń. Wymagałoby to wykorzystania np. algorytmu detekcji mowy (*Voice Activity Detection* – VAD) (Makowski and Hossa, 2020) oraz odpowiedniego mechanizmu separacji mowy od zakłócenia, ponieważ założeniem opracowanego w ramach rozprawy doktorskiej systemu jest obecność w procesie mowy czystej, neutralnej. Wydaje się, że jest to jeden z bardziej istotnych kierunków rozwoju tego typu metodologii.

Innym polem zastosowań mogą być niewątpliwie adaptacyjne systemy rozgłoszeniowe, które zwykle działają w obecności hałasu. System adaptacyjny wykrywający rodzaj hałasu i modyfikujący sygnał mowy na mowę lombardzką może poprawić zrozumiałość mowy w takich warunkach.

Ostatnio pojawiły się również nowe trendy w konwersji głosu, zwane konwersją głosu „wiele-do-wielu” (ang. *many-to-many*) (Lee *i in*., 2021; Luong, Tran, 2021; Merritt *i in*., 2022) oparte na głębokich modelach. W szczególności autoenkodery wariacyjne (VAE) służą do

wyodrębniania cech mówcy i treści językowej z wypowiedzi. Metoda zakłada, że model głęboki może przekształcać głosy z wielu plików źródłowych do wielu źródeł docelowych za pomocą jednego VAE (Luong, Tran, 2021). Można więc założyć, że podejście oparte na VAE może być nowym kierunkiem wykorzystywanym do konwersji mowy naturalnej na mowę lombardzką w warunkach hałasu.

POLITECHNIKA GDAŃSKA

## Acknowledgments

I would like to thank many people for their support and help during the time this dissertation has been written.

Prof. Bożena Kostek – I could always count on your advice and help; you have never left me alone with challenges. Even if I sent the results of the experiments late at night, you always found time to review them and comment on them.

Dr. Grazina Korvel – your ability to put complex things into words was amazing, and I learned a lot from you. I have, however, still a lot to learn from your readiness to help at any moment.

Marcin Dembowski – our small talk led to the concept of this dissertation, and you always motivated me to go ahead.

PGS Software and – especially – Grzegorz Widziszowski, for your full support and for giving me the ability to work on this dissertation.

Witold Bołt – I remember when we first met, and after our discussion, I was sure I wanted to go that way and experiment. Your optimistic perspective was strongly motivating.

All my friends and colleagues – you always supported my ambitions – that was something very important for me.

Finally, I want to thank my beloved wife Dorota and my fantastic sons – Damian and Bartosz. They always supported me, and – although they probably suffered because of a lack of time for them – they were patient and dedicated during this time.

POLITECHNIKA GDAŃSKA

## List of Figures

## List of Tables

# List of Abbreviations and Symbols

## Abbreviations

| | |
|---|---|
| **ANN** | Artificial Neural Network |
| **AO-L** | Audio-only case with the Lombard effect |
| **AUC** | Area Under Curve |
| **CMOS** | Comparative Mean Opinion Score |
| **CNN** | Convolutional Neural Network |
| **D4C** | Definitive Decomposition Derived Dirt-Cheap (band-aperiodicity estimator) |
| **DCT** | Discrete Cosine Transformation |
| **ESTOI** | Extended Short-Time Objective Intelligibility |
| **FFT** | Fast Fourier Transform |
| **GPC** | Gaussian Process Classifier |
| **IF** | Instantaneous Frequency |
| **ITU** | International Telecommunication Union |
| **ITU-R** | ITU Radiocommunication Sector |
| **ITU-T** | ITU Telecommunication Standardization Sector |
| **L-BFGS** | Limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm |
| **LHUC** | Learning Hidden Unit Contribution |
| **LPC** | Linear Predictive Coding |
| **LS** | Lombard speech |
| **LTI** | Linear time-invariant system |
| **MFCC** | Mel-frequency Cepstral Coefficient |
| **MLM** | Machine Learning Model |
| **MLP** | Multi-layer Perceptron |
| **MOS** | Mean Opinion Score |

| | |
|---|---|
| **MOS-LQO** | Mean Opinion Score - Listening Quality Subjective |
| **MUSHRA** | MUltiple Stimuli with Hidden Reference and Anchor |
| **PESQ** | Perceptual Evaluation of Speech Quality |
| **PSOLA** | Pitch Synchronous Overlap and Add |
| **PSQM** | Perceptual Speech Quality Measure |
| **QDA** | Quadratic Discriminant Analysis |
| **ReLU** | Rectified Linear Unit |
| **RMS** | Root Mean Square |
| **ROC** | Receiver Operating Characteristic |
| **SD** | Standard Deviation |
| **SISO** | Single-Input and Single-Output |
| **SNR** | Signal-to-Noise Ratio |
| **STFT** | Short-time Fourier Transform |
| **SVM** | Support Vector Machine |
| **TTS** | Text-to-Speech |
| **VAD** | Voice Activity Detector |
| **VAE** | Variational autoencoder |
| **VOIP** | Voice over Internet Protocol |

**Symbols**

| | |
|---|---|
| $Mel(f)$ | Mel scale |
| $f$ | Frequency [Hz] |
| $c_n$ | MFCC coefficient |
| $h_k(n)$ or $h(t)$ | Impulse response |
| $a_k$ | Amplitude |
| $\varphi_k$ | Phase |

| | |
|---|---|
| $[.]^T$ | Matrix transpose operation |
| $Dist$ | Distance |
| $SD_i\ (lomb)$ | Standard deviation of $i^{th}$ Lombard speech parameter |
| $\underline{r_i}$ | Mean value of the i$^{th}$ vector |
| $max\ \{\}$ | Maximum value |
| $F_0$ | Fundamental frequency |
| $sc$ | Scaling coefficient |
| $\hat{x}(n)$ | Predicted signal |
| $F_i$ | $i^{th}$ formant |
| $Acc_i$ | Accuracy |
| $SBW$ | Spectral bandwidth |
| $SF$ | Spectral flatness |
| $PX(n)$ | Power spectrum |
| $x_i$ | Vector |
| $Y$ | Matrix |

POLITECHNIKA GDAŃSKA

# Table of Contents

# 1 Introduction

The background of this dissertation is connected to communication between people and their interaction with information. The modern world is full of information. To be more precise, the world was always full of information, but modern technology allows for producing and storing gigabytes of data every second. The amount of data stored worldwide − presented by STATISTA (Statista, 2022) – is shown in Figure 1.1.

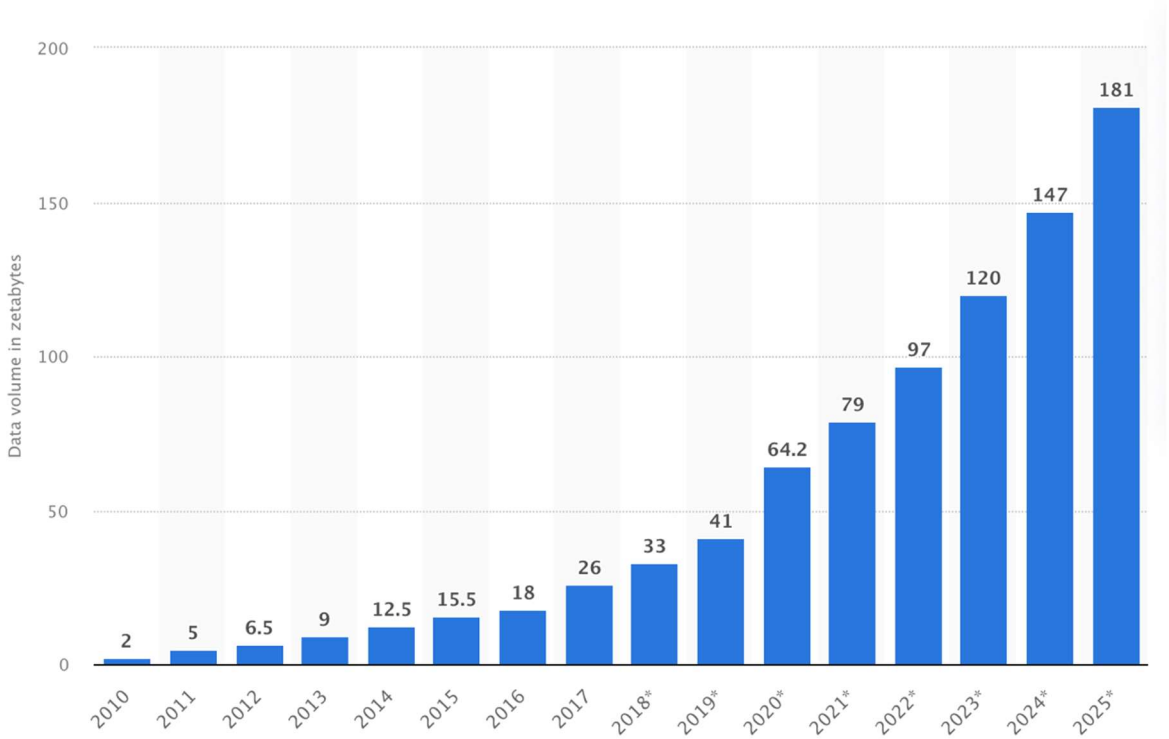



*Figure 1.1 Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 (Statista, 2022)*

The world becomes more "dense" – full of information, as well as full of interactions and communication, even though they are not always direct. It is also evident that technology has changed the world. But technological revolution, digital in its nature, led to the situation where communication is nowadays much more accessible. Moreover, technology is one of the most critical drivers but, at the same time, a valuable asset. There are digital media, mobile phones, the internet with its social applications, and many more means of communication. From one point of view, it simplifies everything. One can quickly write a message to virtually anyone, locate a person, or make a video call with a team or friends. Still, from another perspective – massive communication increases the world's noise.

This noise might be virtual – caused by hundreds or thousands of people on social platforms. But it might be genuine – for example, when many people talk via their mobile phones on a train. Such a real noise can be bothersome. What is more, this overwhelming noise may come from many sources – not only digital ones. But because the world has become "dense," one can observe the overall increasing noise: more cars than ever, more people in big cities using shared space, higher buildings with more and more offices or apartments, massive commercial centers with hundreds of shops in one building – these are just examples of the "dense" world, a world filled with noise.

But again – communication keeps being most important. People live close, work in teams, involving personal or interpersonal communication, so they must communicate. Most of the problems observed in companies come, for instance, from miscommunication – people are not entirely transparent, employees hide their problems, and members of the team do not speak with each other. But how to handle clear communication if noise is everywhere?

This question becomes more and more valid in the contemporary world. There are many situations where speech – as the primary mean of communication - becomes simply unintelligible: speaking in a busy street, listening to messages at the train station, small talk in the disco. These situations are recalled here as examples, but one can easily imagine to what extent speech intelligibility is important and how it becomes degraded in such environments.

It is clear that these two worlds – seemingly opposite – can be joined: technology adding a lot to the world noise and the need for clear communication, which is often destroyed by the technology and its noise. However, technology, on the other hand, might be used for improving intelligibility and thus the ability to communicate.

In other words, the goal is simple: if one wants to understand speech, this speech will have to be heard better than the noise. This goal cannot be easily achieved by the simple way of increasing the volume of the speech. It is easy to imagine how that may impact the world's noise, not to mention the overall speech quality after the primitive operation of increasing volume. That is not the solution sought.

It should be mentioned that nature equipped humans (but not exclusively as other 'species' experience the same effect) with an excellent mechanism of speaking "over the noise," which is called the Lombard effect (Brumm and Zollinger, 2011; Lombard, 1911; Zollinger and Brumm, 2011). This effect occurs when the speaker unconsciously regulates vocal output by changing the volume but also certain acoustic features of the uttered speech in noise. This type of speaking is a natural way of adapting speech to make it more intelligible, especially with the assisted ambient noise (Brumm and Zollinger, 2011; Lombard, 1911; Zollinger and Brumm, 2011). There are several characteristics of speech that are changed when spoken in the presence of background

noise – for instance, spectral tilt, the value of the fundamental frequency, or formants positions. They are to be discussed in detail in Chapter 2.

Then the question is raised: why not use the Lombard effect in the sound transformation to increase the speech intelligibility without increasing the overall sound level? This idea seems valid since the best way of applying technology is to mimic the natural effects by its means.

Therefore, the motivation behind and the aim of this dissertation is to build a system that mimics the natural way of speaking when talking in ambient noise. There is, however, one more important issue to mention. There are papers describing methods of adding the Lombard effect to artificially generated speech, for instance, in text-to-speech (TTS) systems (Bollepalli *et al.*, 2019). But real life is much more complicated – usually, the communication (or messaging) takes place in a "real" scenery with people speaking at the same time – so there is no possibility to transform the speech from "text–to–speech," simply because there is no text. So, the idea is to convert the speech signal itself – and not handle its acoustic, pronunciation, and language model as well as semantic content. Such an approach may be called "speech-to-speech" transformation. The topic formulated this way is much less examined in the literature, especially in the context of the Lombard effect.

This type of sound transformation could be used in multiple applications, e.g.:

- Public address systems – when the people speaking to microphones cannot "feel" the street or other type of noise (e.g., fire/hazard warning alarms); therefore, they are not modifying the way of speaking.
- Hearing aids – Lombard effect could improve the speech intelligibility in noise.
- Personal communication devices – such as mobile phones, when using them in the background noise.

The number of applications could easily be increased – speech intelligibility is one of the most crucial factors in the modern communicative world.

There is also another critical issue to define – how can one measure the efficiency of such a system that improves speech intelligibility. Noises are different – everyone is intuitively aware that speaking over babble speech might also be different from speaking over the street noise. The conditions can vary – noise levels can be high or low – and the effort to talk over the noise may also be different. Moreover, the type of unconscious speech characteristics changes depends strongly on the above conditions and also – which has been examined – on the speaker's gender, or – more precisely – on the vocal speech signal features since a woman's voice might be as low as the man's.

The efficiency should be measured considering these various conditions. Generally, two ways of the efficiency evaluation may be discerned: either using objective metrics (ITU-T Recommendation P.563, 2004; ITU-T Recommendation P.862, 2001; ITU-T Recommendation P.862.1, 2003) or subjective listening tests (ITU-T Recommendation BS.1534-1, 2003). Particularly, there is a possibility to use objective algorithms to determine the accurate value of the speech intelligibility or – more precisely – speech quality, which is highly correlated with intelligibility. There are many objective metrics, but not all apply to the speech intelligibility measurement concerning the Lombard effect. However, in this thesis work, two of them are to be examined: double-ended PESQ (ITU-T Recommendation P.862, 2001) and single-ended P.563 (ITU-T Recommendation P.563, 2004). Both are used and considered as an objective measure of the efficiency of the system proposed.

Because these measures are treated as final system quality metrics, there is one disclaimer here: both metrics are used for verifying the quality of telecommunication channels, and thus they are not fully adequate for measuring speech in reverberant environments (Odya *et al*., 2021). It should, however, be mentioned that they are still applied to such a situation. However, the final verification in the dissertation does not consider the possibility of speech convoluted with reverberation. In addition, the system must abstract from the predefined set of conditions and noises − it should cover an extended scope of possible environments in real- or nearly real-time.

Subjective metrics can be measured using many different methodologies. Still, this work is focused on using the MUSHRA test (ITU-T Recommendation BS.1534-1, 2003), which seems to provide meaningful results when testing speech quality. However, the drawback of using subjective examinations is the need to limit the number of tested recordings, which may effectively lower the credibility of the results.

Taking the above into consideration, there is a need to define the overall aim of this work, which is to create a system that can improve speech intelligibility in various noise conditions depending on the type of speaker.

This, however, brings another question – what if the speech uttered is already “Lombard-type”? There is no need to modify that kind of speech since it might lead to quality degradation rather than improvement. Therefore, one part of the system is devoted to Lombard speech detection – a component responsible for avoiding modifying speech that has already been changed naturally by the speaker.

The other component should also be responsible for detecting noise characteristics – to allow for proper speech modifications relevant to the given conditions. This is a necessary component to enable adaptive transformation, based on noise profiling as called in this thesis.

Assuming that the system is aware of the speech and noise types, the predefined improvements to the speech signal might then be applied. Such an approach based on "if-this-then-that" logic suggests that the system might be efficient but not flexible and not adapt to new conditions. What is more, this approach is not generalizing well and can − if treated as a type of "machine learning" − be easily overfitted. Therefore, in this work, machine learning – based on a neural networks approach to adaptive speech modifications is presented.

This dissertation has three main theses:

1. **Employing a Convolutional Neural Network (CNN) enables the detection of the Lombard effect in uttered speech with sufficient accuracy for the purpose of speech profiling**
2. **Baseline machine learning methods can be used to effectively profile the ambient noise in a stable and near real-time manner.**
3. **It is possible to build a system that adaptively modifies the speech signal to improve the speech intelligibility in noise optimally based on speech and noise profiling.**

This dissertation is structured in a way that follows the theses presented:

1. First, the Lombard speech is defined, and speech detection methods are described.
2. Then the potential speech modification methods are presented with their theoretical background. The Lombard effect generation methods are recalled, and state-of-the-art methods are briefly described.
3. Next, the experiments performed are individually reported for all parts of the system: speech type detection, noise profiling, and adaptive speech modification. The system is then completed, validated, and compared with the typical speech transformations.
4. Finally, conclusions are derived along with scientific contributions and plans for future research.

The overall structure of the system is presented in Figure 1.2.

It should be noted that several expressions have the same meaning concerning the experiments carried out within this Ph.D. work, namely, speech/noise detection, speech/noise profiling, and Lombard/non-Lombard speech recognition/classification. Even though their meaning is not precisely the same, they are used interchangeably.

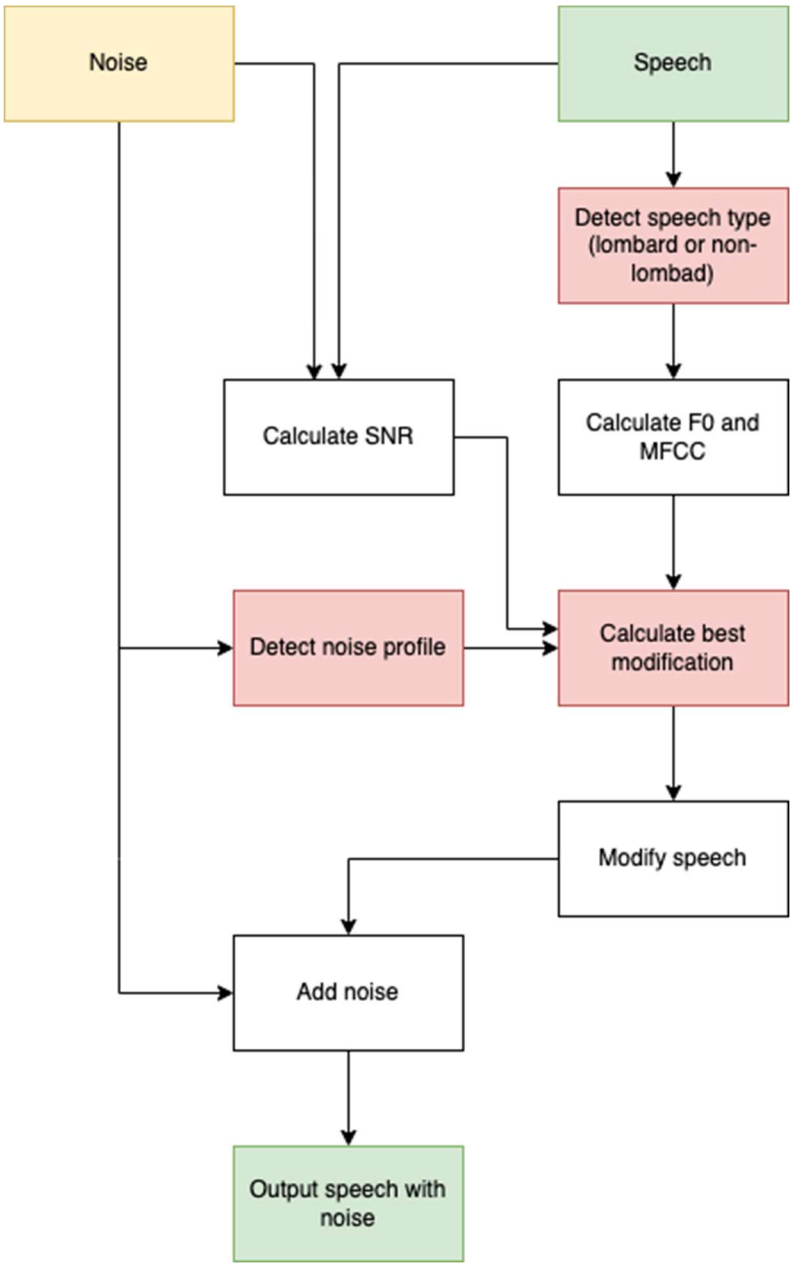



*Figure 1.2 Adaptive speech modification system*

The system validation is performed using the following processes presented in Figure 1.3:

1. First, the clean speech is profiled to determine whether the Lombard effect is present or not.
2. Then the input noise is profiled.
3. Using the above results, the adaptive system is implemented, and speech is modified accordingly.
4. Then noise is appended to the modified speech, and P.563 (singled-ended metric) is measured.

5. Moreover, ambient noise is added to the clean speech, and P.563 is calculated.
6. As the last step, both calculated P.563 metrics (for clean and processed speech) are compared.

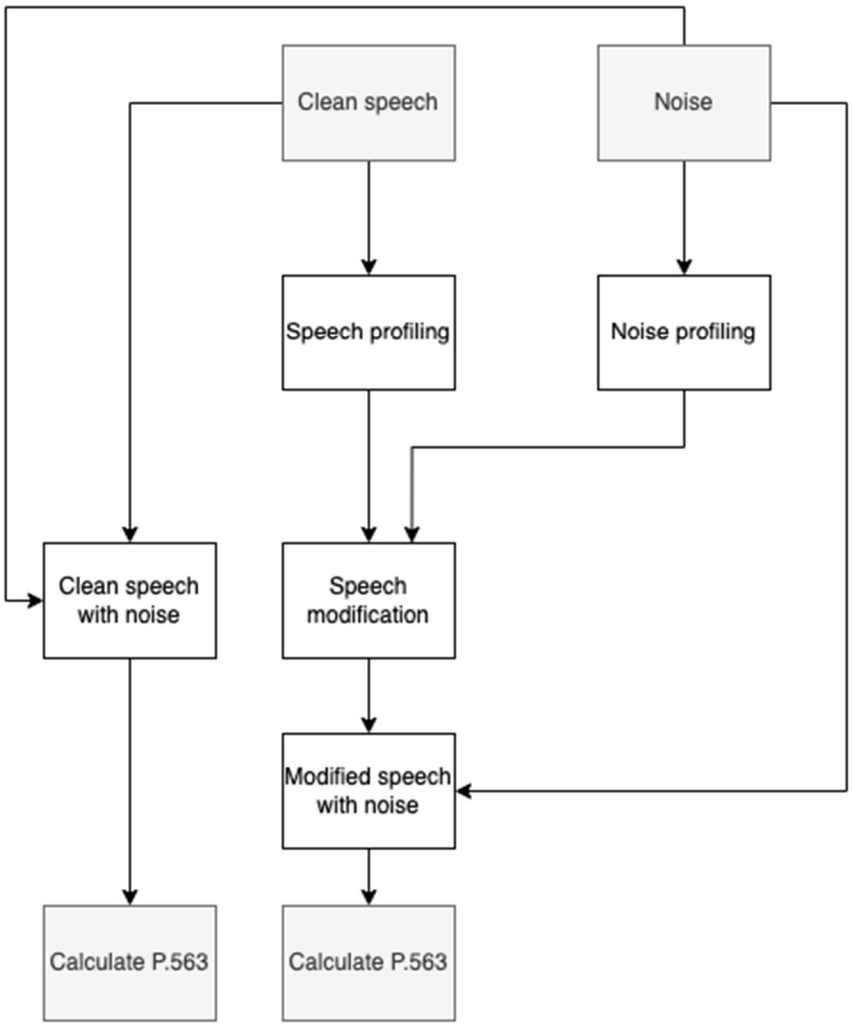



*Figure 1.3 Process of the system efficiency verification*

It seems important to recall the methodology that is to be employed at particular stages of this dissertation realization (after experimental testing and verification):

1. For Speech profiling (i.e., Lombard speech detection) Convolutional Neural Network (CNN) with several 2D signal representations is used. CNNs – although designed to work with image recognition challenges – are also widely used in sound classification problems (Mu *et al.*, 2021).

2. For Noise profiling typical, a simple Naïve Bayes classifier is employed. It is fast in both learning and inference, and the classification task has successfully been solved with this simple model.
3. Adaptive speech modifications are implemented using two components:
    a. Finding the best modification parameters for a signal based on the pre-trained MLP (multi-layer perceptron) classifier.
    b. Modifying the signal using WORLD vocoder and signal processing algorithms to change both fundamental frequency (F0) and formant values.

It is important to underline the fact that the experiments visible in the literature (Bollepalli *et al.*, 2019) are implemented using text-to-speech (TTS), which limits their applications in real-time systems, processing the speech signal recorded directly.

Finally, the overall dissertation structure is presented in Figure 1.4. The components described in the given chapters will have a grey background displayed in the blocks. For example, the block referring to "Noise characteristics" is highlighted in grey in Figure 1.4.

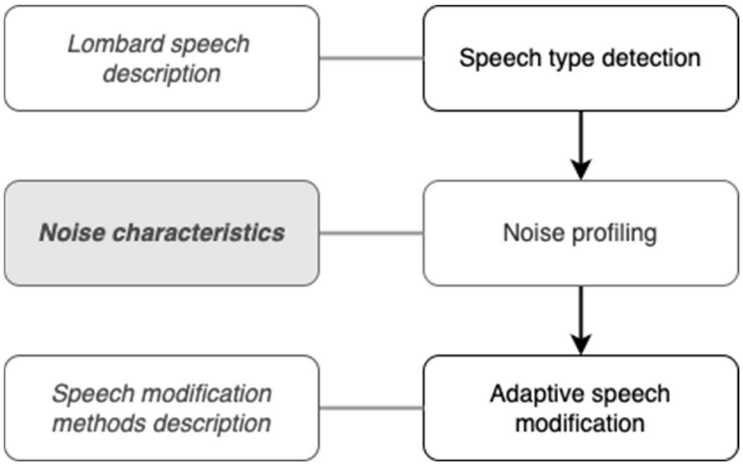



*Figure 1.4 Structure of the dissertation*

# 2 Theoretical background

In this Chapter, the background of the Lombard effect as well as signal analysis applied to detect this effect in speech is presented. In Figure 2.1, the blocks referring to these issues are highlighted in grey. Moreover, background on noise profiling is also presented.

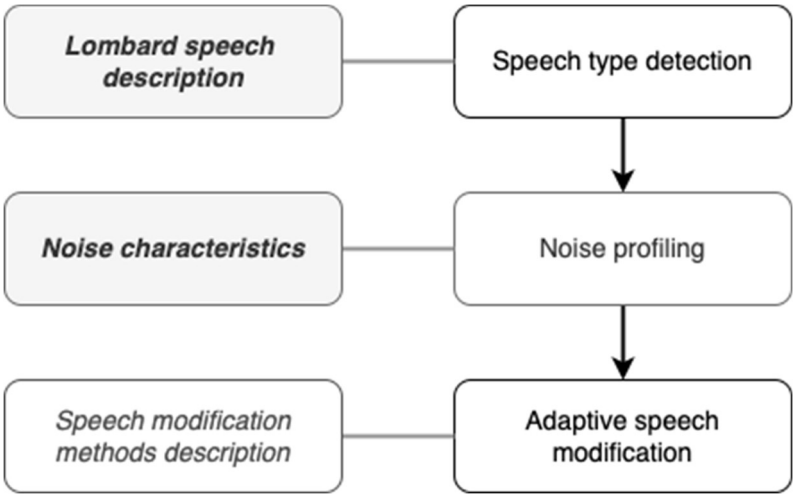



*Figure 2.1 Structure of the dissertation with current Chapter topics highlighted*

## 2.1 Lombard effect background

The Lombard speech is an effect discovered in 1909 by Etienne Lombard, a French otolaryngologist (Lombard, 1911). This effect occurs when the speaker unconsciously changes certain acoustic features of uttered speech in noise. Lombard observed that, for example, speakers in the presence of the crowd speak slightly differently from when they have the opportunity to talk in a more intimate situation.

Numerous studies on the Lombard language (Bořil *et al.*, 2007; Egan, 1972; Kleczkowski *et al.*, 2017; Lu and Cooke, 2008; Stowe and Golob, 2013; Therrien *et al.*, 2012; Vlaj and Kacic, 2011; Zollinger and Brumm, 2011) have identified many features characteristic to this type of expression (Bapineedu, 2010; Bořil *et al.*, 2006; Junqua *et al.*, 1999; Kleczkowski *et al.*, 2017; Lau, 2008). First of all, they include an increase of the fundamental frequency rising or shifting energy from lower frequency bands to medium and higher frequencies. However, features of the Lombard speech are diverse and include other phenomena:

- increasing the level of sound intensity,

- the fundamental frequency rising,
- shifting energy from lower frequency bands to higher frequency bands,
- increase in the value of formants, mainly F1 and F2,
- increasing the duration of vowels,
- increasing the spectral tilt (spectral tilt), etc.

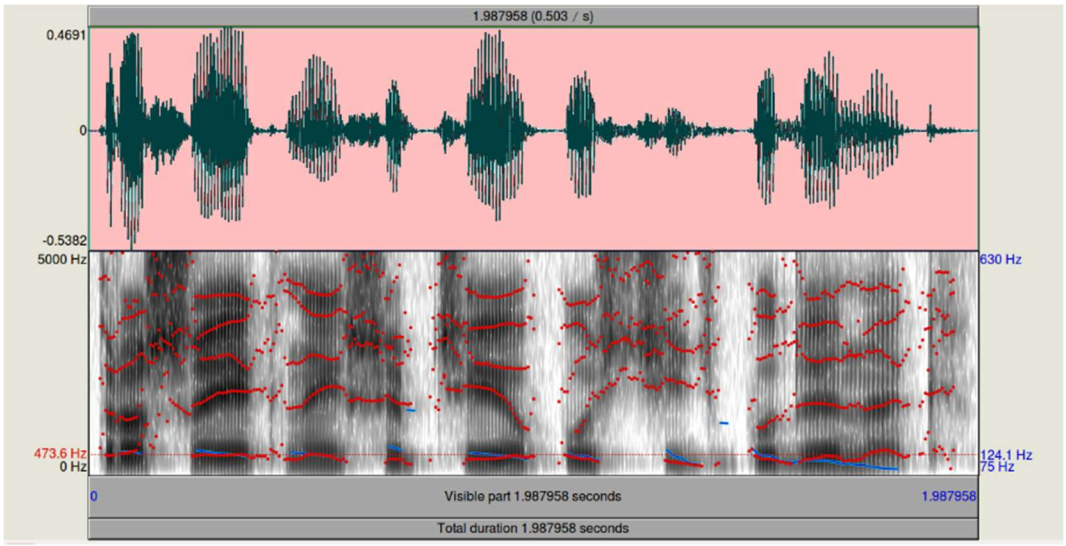
Most of these features can easily be determined but observing changes in these features in the context of Lombard speech is not so simple. The measurement of the instantaneous value of the frequency of the formants is not reliable in such a case because it may indicate, for example, a temporary change associated with the emotions contained in the statement. Regarding Lombard speech detection, it is easier to use long-term measures, e.g., median values of the fundamental frequency should be observed in such a way.

Moreover, it was also reported that this effect, even though involuntary, may be inhibited and trained in the presence of noise (Brumm and Zollinger, 2011; Pick *et al.*, 1989; Therrien *et al.*, 2012). The discussion between spontaneous and the inhibited Lombard effect is still ongoing (Brumm and Zollinger, 2011; Patel and Schell, 2008). Since the discovery was related to the audiology domain, it is not surprising that the first applications were related to speech-in-noise audiometry. The interest in employing the Lombard effect in the medical domain also led to improving low voice intensity in Parkinson's disease patients (Adams and Lang, 1992; Stathopoulos *et al.*, 2014), even though applying elevated noise levels in humans for everyday communication seems a challenging concept to be fully approved. Most of both research and application areas are, however, related to human (and human-computer) communication, telecommunications, etc. (Bapineedu, 2010; Jokinen *et al.*, 2014; Marxer *et al.*, 2018). Especially important are strategies for improving speech comprehensibility in noisy conditions based on various techniques, including speech modeling.

As mentioned earlier, several features of Lombard speech have been identified in numerous studies, including raising the fundamental frequency or shifting energy from lower frequency bands to medium and higher frequencies.

In Figure 2.2, an example of speech signal waveforms along with spectrograms of a sentence uttered in silent and noise (babble speech) conditions is shown. The increase in level and duration of the utterance are visible; other effects are discerned only after quantitative analysis.

a)




b)

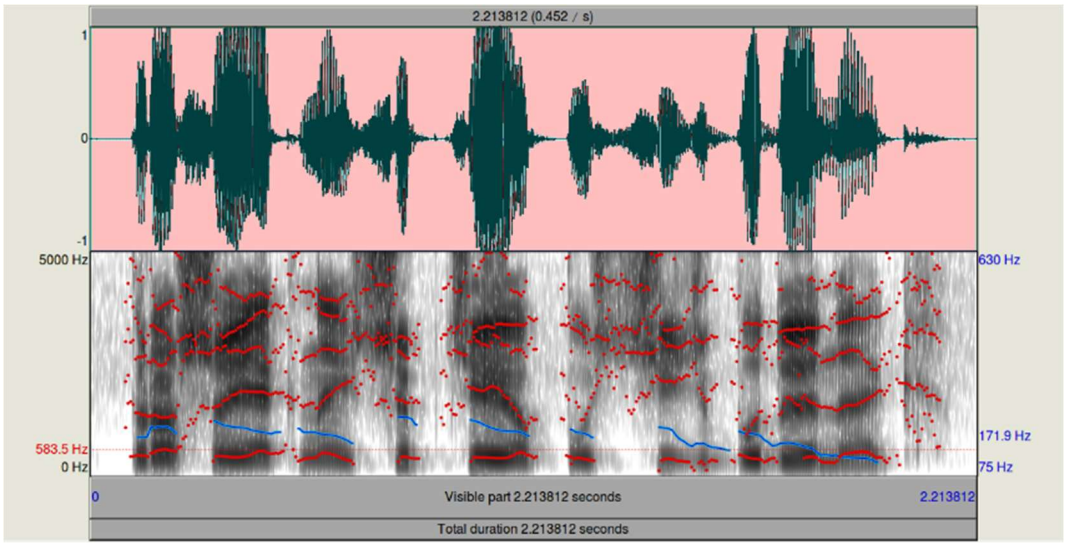



*Figure 2.2 Spectrogram of a sentence in Polish: original sample, silent recording conditions (a), Lombard speech occurrence (an increase of level and duration change are visible) (b)*

In Figures 2.3-2.6 spectrograms and pitch diagrams are presented. Figures 2.3 and 2.5 show spectrograms of the female and male voices, respectively. The top image in each figure presents the Lombard speech, and the bottom image is the neutral speech of the same utterance. It might be observed that the neutral speech's energy is more concentrated in the lower regions, while the Lombard speech energy is moved to higher frequencies and more spread across the frequency range.

Figures 2.4 and 2.6 present the pitch diagrams of the same utterances (female and male voices, respectively) used for the spectrograms. It is clear that the pitch in Lombard speech is much higher than the one of the neutral speech (average pitch is presented on the left-hand side of the diagram).

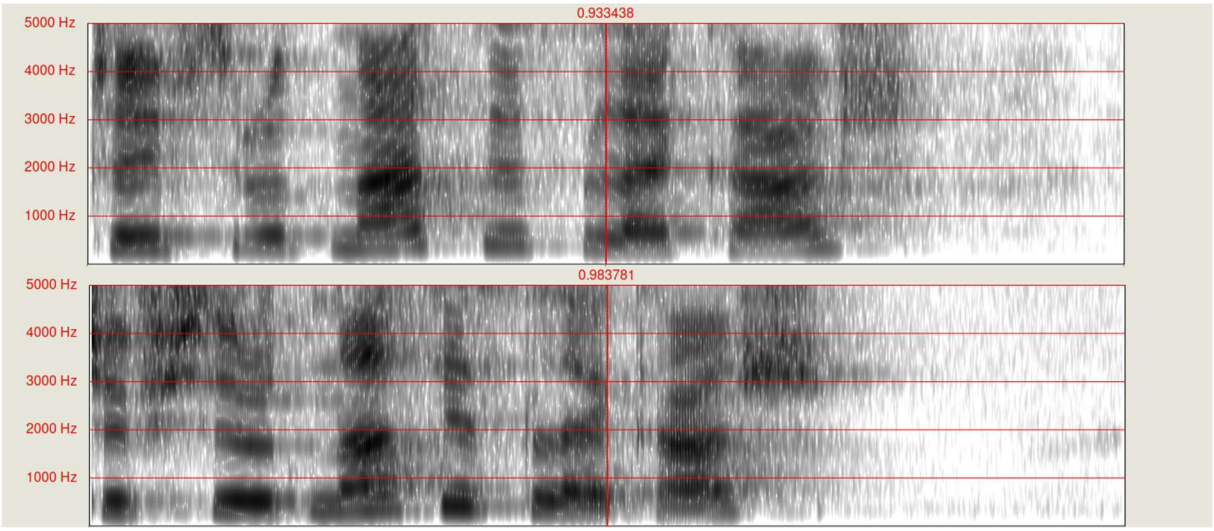



*Figure 2.3 Spectrogram of the female voice recording. The top picture - Lombard speech, bottom - neutral (non-Lombard) speech*

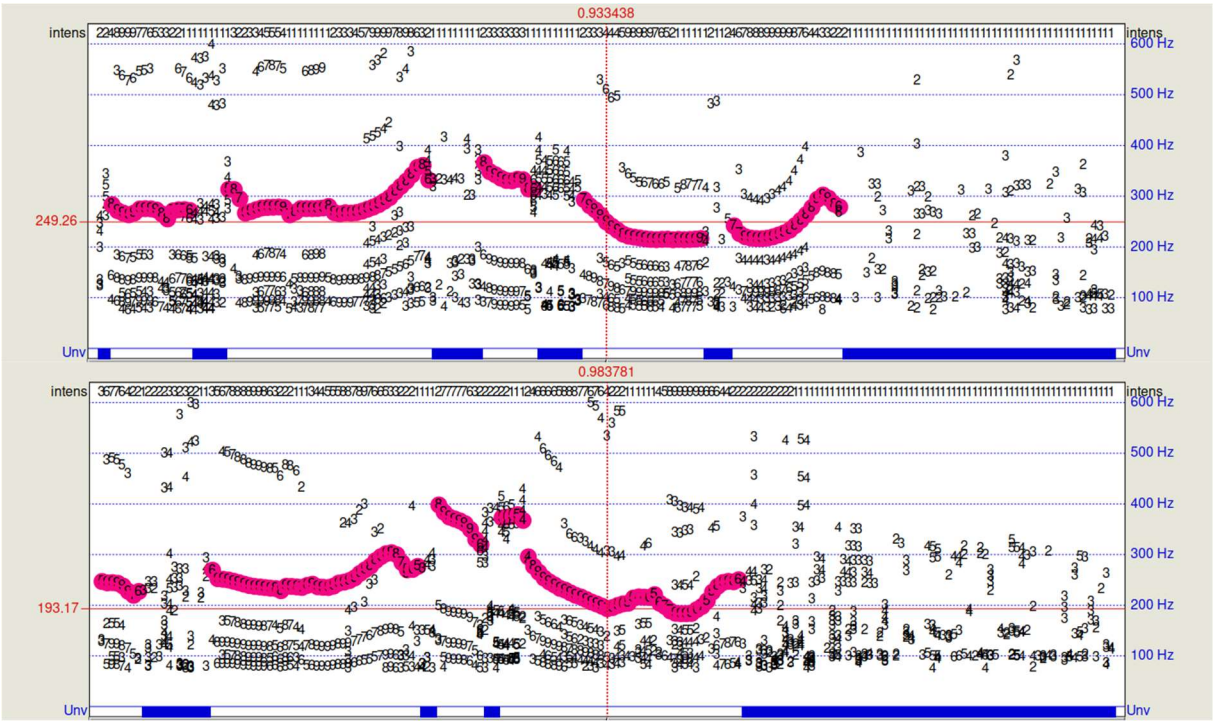



*Figure 2.4 Pitch diagram (F0 marked red) of the female voice recording. The top picture - Lombard speech, bottom - neutral (non-Lombard) speech*

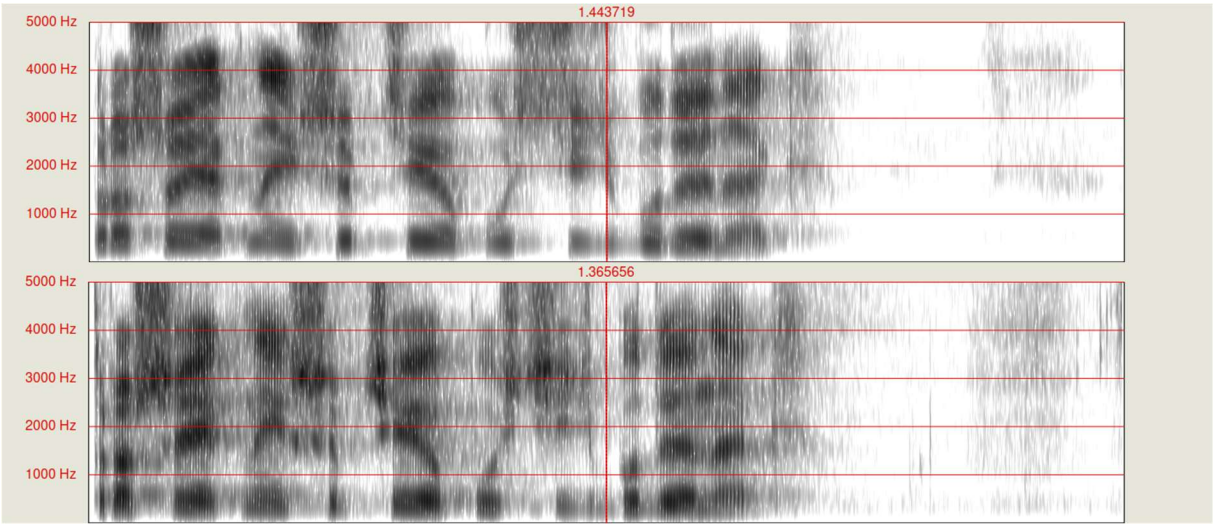



*Figure 2.5 Spectrogram of the male voice recording. The top picture - Lombard speech, bottom - neutral (non-Lombard) speech*

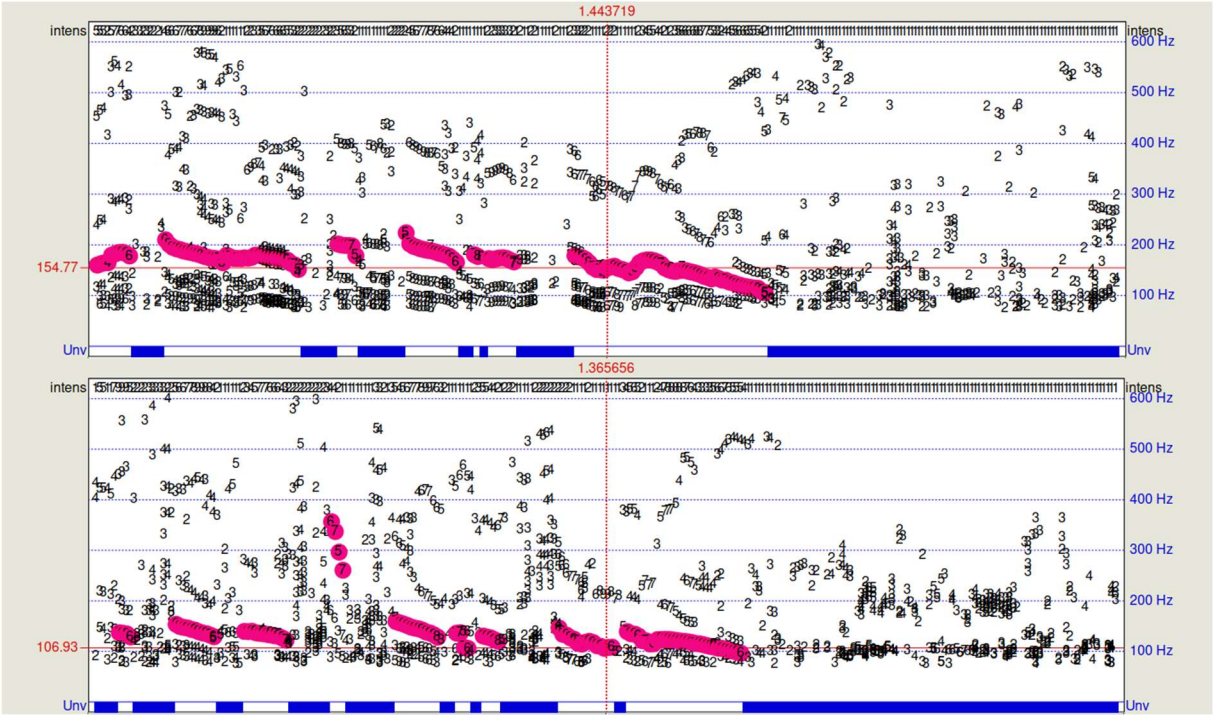



*Figure 2.6 Pitch diagram (F0 marked red) of the male voice recording. The top picture - Lombard speech, bottom - neutral (non-Lombard) speech*

## 2.2 Machine learning approach applied to the Lombard effect detection

In this Chapter, a machine learning background is given for the research on Lombard speech detection. Generally, it may be implemented using different approaches (Bishop, 2006; Goodfellow *et al.*, 2016; Duch *et al.*, 2000; Rutkowski *et al*., 2021).

The simplest one concerns a typical signal-based method (Benesty *et al*., 2008; Tadeusiewicz, 1998): the fundamental frequency value can be calculated and then verified if it is above some – empirically calculated – threshold level. Apparently, raising fundamental frequency value is one of the most significant attributes of Lombard speech.

This method, however, has some limitations:

- Without any reference recording, it cannot easily be defined what level of F0 can be considered as risen.
- The fundamental frequency differs for male and female voices; therefore, it might even be impossible to define any decision level.
- To make this method work, it must be possible to determine the speaker's gender first, which requires a machine learning approach.

### 2.2.1 Lombard speech detection using CNN

Therefore, another approach has been proposed and adopted in this work, namely Convolutional Neural Networks (CNNs). CNN is a regularized multilayer perceptron, and it takes advantage of the hierarchical data, applying convolutional filters, which parameters are to be defined in the process of learning.

CNN is a type of neural network designed to maintain the spatial integrity of the processed image. Typical ANN (Artificial Neural Network) can be used to work with 2D representations by transforming their pixel values into a 1-dimensional vector and then using this vector as input to the first layer of a deep neural network. This might work; however, such transformation means that the spatial relations between pixels are lost.

CNN treats 2D representations differently: it extracts features from the processed image by sliding a convolutional filter over an image and calculating the feature maps. These feature maps create a set of new pixel values calculated from the source image and filter. Convolutional filters applied to the source image have the following attributes:

- Size of the filter tensor (it is usually a 2- or 3-dimensional since the image might be greyscale or color);

- Stride – the number of pixels by which the filter is moved in the subsequent steps;
- Padding – whether the resulting pixel set should be padded with empty pixels to retain the exact size of the feature map as the source image;
- How many filters should be applied to the image.

As with the typical ANN, CNN filtering is done in layers, and the calculated pixel values of the feature maps are then passed to the next layer, using the activation function (usually ReLU in CNN). The next layer works in the same way – so the feature maps are then calculated from the existing feature maps, and so on. The goal of the training process in Convolutional Neural Networks is to get the correct filter values to obtain the best feature extraction characteristics (Goodfellow *et al.*, 2016).

There is also one operation that might be implemented between layers – pooling. It reduces the size of the processed feature map to be processed in the next layer. There are several options here:

- Max pooling – selecting the maximum pixel value of a given pooling area;
- Sum pooling – summing the pixel values within the pooling area;
- Average pooling – calculating the average value of a pixel within the pooling area.

Usually, after the convolutional layers, there is a dense (flat) layer consisting of a number of typical ANN's neurons, which allows using the classification approach by setting up the output neurons' number equal to the number of predicted classes.

Convolutional neural networks conceptually work in the following way:

- Maintain the spatial integrity of the image pixels;
- Extract features of the source images by filtering the image with the set of filters;
- Enhance the feature maps by using the ReLU activation function;
- Reduce dimensionality by using pooling;
- Classify using the dense layers.

A typical block diagram of the convolutional neural network is presented in Figure 2.7.

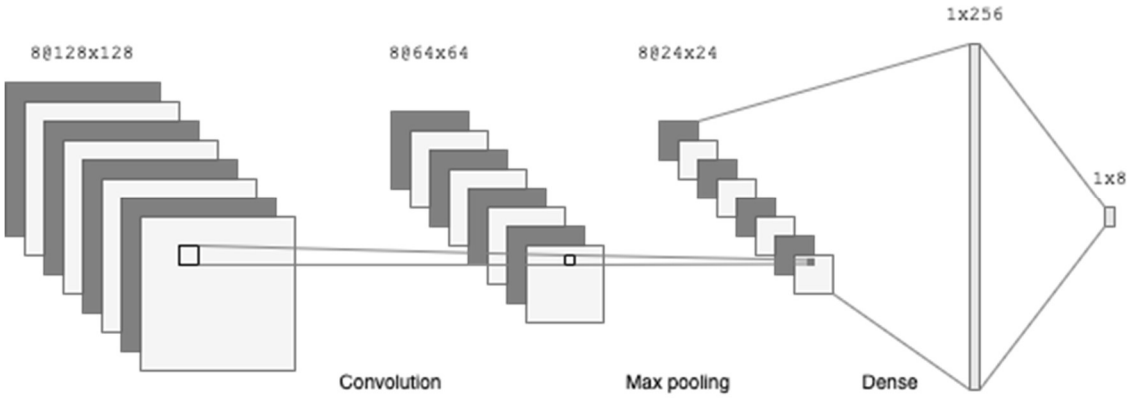



*Figure 2.7 Typical CNN network with three convolutional layers with max-pooling and one dense layer and classification output*

Each unit in the structure of CNN receives input from other units in its neighborhood. It means that the network focuses on local data changes and allows for simple detection of edges, contrasting areas, but also similar features in speech spectrograms (LeCun *et al.*, 1999).

CNN, to work correctly with an audio signal, requires a 2D representation at its input. Sample speech 2D signal visualizations are presented in Figure 2.8, and they are shortly described in the following chapters.

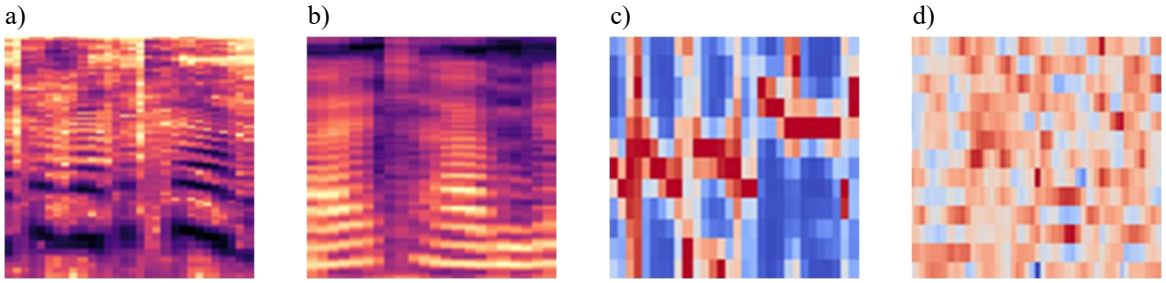



*Figure 2.8 Sample speech signal visualizations. a) Spectrogram, b) Mel spectrogram, c) Chromagram, d) MFCC-gram*

### 2.2.2 Spectrogram visualization

A spectrogram is a visual 2D representation of the signal energy distribution in frequency and time domains.

Let:

$$x = [x(1), x(2), .., x(N)]^T \tag{2.1}$$

be a sequence of samples of the analyzed speech signal, where $N$ is the number of samples per signal and the $T$ superscript placed on the matrix (i.e., $[.]^T$) refers to the matrix transpose operation.

The spectrogram construction process is based on calculating the Short-time Fourier Transform (STFT) for this speech signal. The magnitude spectrum of the $l$-th short-time segment (denoted by $X_l$) is obtained by the following formula:

$$|X_l(k)| = \frac{1}{M_{FT}} \sqrt{(X_l(k))_{re}^2 + (X_l(k))_{im}^2} \tag{2.2}$$

where $X_l(k)$ is the Fourier transform of the short-time segment $x_l$, $k = 1, \dots, M_{FT}$ ($M_{FT}$ refers to the number of Fourier transform coefficients), $l = 1, \dots, L$ ($L$ refers to the number of short-time segments).

### 2.2.3 Mel spectrogram visualization

Mel spectrogram is a mel-scaled power spectrogram. For this purpose, the mel filter bank is constructed over the frequency range from the lower to the upper frequency. The mel spectrum is obtained by multiplying the spectrum coefficients by the filter coefficients. The relationship between the mel scale and the Hertz scale can be described by the following formula:

$$Mel(f) = 2595 \cdot log\left(1 + \frac{f}{700}\right) \tag{2.3}$$

where $f$ is a given frequency in Hertz. Mel scale is fundamental in applications of speech processing because it reflects human perception of sound.

### 2.2.4 Chromagram visualization

Chromagram is another type of representation (and visualization) where the entire spectrum is projected onto 12 bins representing the 12 semitones of the musical octave.

As discussed by Müller (Müller, 2015), the human perception of the pitch has a “color” periodicity, which means that two pitches are perceived similar (in their “harmonic role”) if they differ by an octave. This resulted in an observation that every pitch might be represented by two factors: tone height and chroma. Tone height is represented by the octave number, while the chroma is the number of pitch inside the octave (0 to 11) – just like sounds in a chromatic scale (C – C# - D – D# - … – B).

Chromagram can be created by summing up all coefficients belonging to the same chroma, and it is derived from a pitch-based log-frequency spectrogram having 127 coefficients (Zalkow and Müller, n.d.).

Due to its "musical" context, it does not fit well with the speech visualization problem.

### 2.2.5 MFCC-gram visualization

Mel-frequency Cepstral Coefficients (MFCCs) are a compressible representation of mel spectrogram. To get MFCCs, a log magnitude of mel-spectrum is calculated, and then Discrete Cosine Transformation (DCT) is applied. The mathematical expression of MFCCs is as follows:

$$c_n = \sum_{i=0}^{M-1} m_i \cos\left(\frac{\pi n(i+\frac{1}{2})}{M}\right) \quad (2.4)$$

where $m_i$ are the log filter bank amplitudes, $M$ is the number of filters, $n = 1, \dots, M-1$.

## 2.3 Noise profiling methodology

In this Chapter, the theoretical background for noise profiling is given. First of all, the algorithms used for noise type recognition are described. Then the spectral characteristics used as an input for the detection models are presented.

### 2.3.1 Baseline algorithms

All classification models employed in the noise profiling task are briefly described in the following subsections.

**Naïve Bayes**

A Naïve Bayes classifier is a simple probabilistic classifier based on the assumption that the predictors (features) are independent (Barber, 2011; Zhang, 2004). That is why it is called "naive." Although it is a simple model, it can provide high accuracy levels.

Bayesian probability means that the a posteriori probability can be calculated using the following formula (Li *et al.*, 2016):

$$P(C_k|\boldsymbol{X}) = \frac{P(C_k)P(\boldsymbol{X}|C_k)}{P(\boldsymbol{X})} \quad (2.5)$$

where **X** represents the vector with n conditionally independent features $X_1, X_2, \dots, X_n$ , and $C_k$ is a possible outcome class.

**Linear SVM**

Support Vector Machines (SVMs) are types of supervised learning models that have associated learning algorithms and are used for both classification and regression tasks (Cortes and Vapnik, 1995; Platt, 1999). Support vector machine constructs a hyperplane (or a set of hyperplanes) that separates data in a way that provides the largest distance to the nearest training data point of any predefined class. It is necessary to select a proper kernel function that suits the given problem (Cortes *et al.*, 2004).

For SVM, a set of training dataset points of the form $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in R^n$, $y_i \in \{1, -1\}$, $i = 1, \ldots, k$ ($k$ is the number of instances) is given. A kernel used to train linear SVM takes the following form:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) \tag{2.6}$$

where $\phi$ is a function that mapped training data into higher dimensional space.

The following parameters of linear SVM were implemented: regularization C=0.025, probability estimates have been enabled, and tolerance for stopping criterion is equal to 0.001.

**SVM with polynomial kernel**

The polynomial kernel used in SVM can be defined as follows (Cooke *et al.*, 2019; Wu *et al.*, 2004):

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma \boldsymbol{x}_i^T \boldsymbol{x}_j + c)^d \tag{2.7}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are vectors of features computed from training or test samples, $\gamma > 0$ and $c \geq 0$ are free parameters trading off the influence of higher-order versus lower-order terms in the polynomial. Finally, $d$ is the degree of the polynomial. In the case discussed, the degree is 3.

The following parameters of the polynomial SVM were implemented: regularization parameter $C = 1$, gamma coefficient ($\gamma$) set to auto (which means that it uses 1/number_features), probability estimates were enabled, independent term in kernel function equals 0, tolerance for stopping criterion is equal to 0.001.

**Gaussian process classifiers**

Gaussian process classifiers (GPCs) use a generalized Gaussian probability distribution (Rasmussen and Williams, 2006). GPCs may be expressed in terms of the kernel models, and they can predict highly calibrated class membership probabilities, but the selection (and configuration) of the kernel is difficult. In the test, the exponential kernel was used − it takes one base kernel (Watanabe *et al.*, 2017) and combines them via:

$$k_{exp}(\boldsymbol{X}, \boldsymbol{Y}) = k(\boldsymbol{X}, \boldsymbol{Y})^p \tag{2.8}$$

where **X** and **Y** are two matrix of datapoints. In this research the exponent is equal to 2.

As a source kernel, a Rational Quadratic kernel was used. It is parameterized by the length scale parameter and a scale mixture parameter. The kernel is given by (Bhavan *et al.*, 2019):

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(1 + \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^2}{2\alpha l^2}\right)^{-\alpha} \tag{2.9}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are vectors of features computed from training or test samples, $\alpha > 0$ is the scale mixture parameter, $l > 0$ is the length scale of the kernel.

The L-BFGS-B (a limited memory Broyden–Fletcher–Goldfarb–Shanno) algorithm, an extension of the L-BFGS algorithm to handle simple bounds on the model Zhu et al. (Byrd *et al.*, 1995; Zhu *et al.*, 1997), serves as a large-scale bound-constrained optimizer algorithm (Byrd *et al.*, 1995; Zhu *et al.*, 1997). It is used in the context of finding a (local) minimum of an objective function.

**Decision Tree**

Decision tree learning is a predictive modeling approach used in machine learning (Kamiński *et al.*, 2018). It uses a decision tree as a predictive model that leads to conclusions. Its goal is to create a model that predicts the value of the target variable based on a set of input features. The parameters used in this test are as follows: the quality of the split is Gini impurity, maximum depth of the tree is 5.

**Random Forest**

Random forests work similarly to decision trees, but it builds an ensemble of decision trees and returns the class that is the statistical model of the classes returned by the tested decision trees (Ho, 1995).

Parameters used in this test: the quality of the split is Gini impurity, the maximum depth of the tree is 5, and the number of estimators (trees in the forest) is set to 10.

**MLP Classifier**

MLP (Multilayer Perceptron) is a type of feedforward neural network that consists of at least three layers - with one hidden layer (in its minimum form). Another essential feature is that MLPs have a non-linear activation function. In the implementation, there is one hidden layer with 100 neurons, and the ReLu activation function is used. The optimizer used for weight is Adam optimization, which refers to the stochastic gradient descent optimizer (Pedregosa *et al.*, 2011).

The following parameters of the MLP classifier were used: L2 regularization parameter (alpha) is set to 1, and the maximum number of iterations equals 1000. The hidden layer contains 100 neurons, and the activation function is ReLU. Adam solver for weight optimization was used.

**AdaBoost classifier**

The AdaBoost classifier is an ensemble-type meta-estimator that begins with fitting the classifier on an original dataset and then performs additional fitting but focuses on complex cases. When the perfect fit is achieved, the learning stops (James, n.d.; Rojas, 2009).

The following parameters were used: the maximum number of estimates at which boosting is stopped equals 50, the learning rate equals 1, and SAMME.R is used as the boosting algorithm.

**Quadratic Discriminant Analysis**

Quadratic Discriminant Analysis is a classifier with a quadratic decision surface (James, n.d.). It can be derived from a simple probabilistic model - they can model the class distribution of the data $P(X|y_k)$ for each class $k$. This is based on the Bayes rule presented above in the description of the Naive Bayes classifier. If there is an assumption that the covariance matrices are diagonal, then the input features are assumed independent - the resulting classifier is then equivalent to Naive Bayes. For the test, the regularization parameter is set to 0.

### 2.3.2 Parameters for noise profiling

**Spectral centroid**

Spectral centroid is a metric used in digital signal processing that characterizes the spectrum of the signal. It allows calculating where the center of mass of the spectrum is located. This measure is related perceptually to the impression of the sound brightness. In this research, the spectral centroid is calculated as the weighted mean of the frequencies present in the signal with their magnitudes as the weights (Li, 2021):

$$SC = \frac{\sum_{n=0}^{N} f(n)X(n)}{\sum_{n=0}^{N} X(n)} \tag{2.10}$$

where $X(n)$ is the weighted magnitude of the Fourier transform at frequency bin $n$, and $f(n)$ represents the center frequency of that bin.

**Spectral bandwidth**

The spectral bandwidth (SBW) is used to define the bandwidth of the spectrum. This measure shows the concentration of spectrum around the centroid and is computed by (Krčadinac *et al.*, 2021):

$$SBW = \left(\sum_{n=0}^{N} X(n)(f(n) - SC)^p\right)^{1/p} \quad (2.11)$$

where $X(n)$ is the weighted magnitude of the Fourier transform at bin $n$, $f(n)$ represents the center frequency of that bin, $SC$ is the spectral centroid (see Eq. (2.10)). Variable $p$ is equal to 2 − this corresponds to a weighted standard deviation around the centroid.

Spectral bandwidth values were calculated for all analyzed noise types and frames within the signal. The statistical features of the calculated values are sufficiently separated. This conclusion is a basis for the hypothesis that statistical parameters of the frequency analyses can provide a sufficient degree of predictive ability to perform noise classification.

**Spectral flatness**

Spectral flatness is a measure of an audio sound spectrum that provides a way to quantify how tone-like a sound is, as opposed to being noise-like. High spectral flatness - approaching 1.0 for white noise - means that the spectrum has a similar amount of power in all spectral bands. Low spectral flatness values (approaching 0.0) convey that the power is concentrated in a small number of bands − typically, it is a mixture of sine waves.

The spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum (Gosztolya, 2019), i.e.:

$$SF = \frac{\left[\prod_{n=0}^{N-1} PX(n)\right]^{1/N}}{\frac{1}{N}\sum_{n=0}^{N-1} PX(n)} \quad (2.12)$$

The power spectrum  at bin number  is given by the following formula (Mencattini *et al.*, 2014):

$$PX(n) = \frac{1}{N}\sqrt{X(n)_{re}^2 + X(n)_{im}^2} \quad (2.13)$$

where $X(n)$ is Fourier transform coefficient at bin $n$, $re$ means a real part, and $im$ – an imaginary part.

# 3 Selected speech modification methods

This Chapter focuses primarily on analyzing speech signals uttered in a simulated noisy environment, then modifying clean speech and artificially mixing it with some noise types. The are several approaches available to this end: PSOLA (Moulines and Charpentier, 1990) and LPC algorithms (O'Shaughnessy, 1988), harmonic models (Korvel *et al.*, 2016, 2019), source-filter models (Kawahara, 2006; Morise *et al.*, 2016) or sinusoidal models (Ellis, 2003), to name a few. Methods based on machine learning algorithms, usually used in text-to-speech applications, are also described.

Analysis of the Lombard speech often is related to subjective assessment of speech intelligibility. There are, however, objective indicators such as PESQ (Perceptual Evaluation of Speech Quality) or P.563, which are used in speech quality studies of telecommunications channels (Beerends *et al.*, 2009, 2013; ITU-T Recommendation P.563, 2004; ITU-T Recommendation P.862, 2001). That is why such measures are also recalled and discussed in this Chapter.

Moreover, a method based on speech signal features was introduced by the author of this dissertation for the assessment purpose.

Fig. 3.1 shows the current chapter topic highlighted in grey.

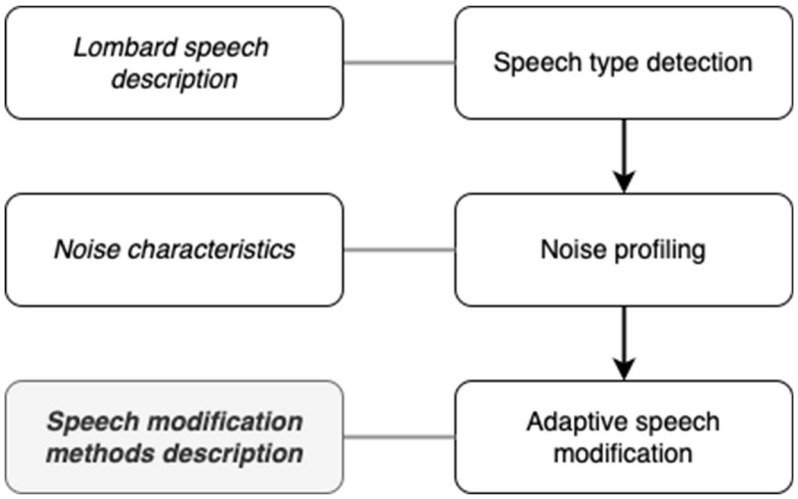



*Figure 3.1 Structure of the dissertation with current Chapter topics highlighted*

## 3.1 Speech signal modifications

**Frequency-domain filtering**

The first modification considered in this work is related to focusing on the specific frequency regions of the speech signal. These regions are obtained by applying a filter in the frequency domain. Two signal filtering options are considered:

- Application of a low-pass filter to the speech signal, which cuts high frequencies.
- Application of a high-pass filter to the speech signal, which cuts low frequencies.

Filtering was applied with the Praat built-in filtering function (Kurban Ubul *et al.*, 2009). This function multiplies the complex spectrum in the frequency domain by a real-valued filter function, which has the symmetric Hann-like band shape. The width of the region between pass and stop equals 100 Hz.

**Manipulation of duration**

Changing speech duration results in speeding up or slowing down the analyzed speech signals. In order to modify the signal duration, the PSOLA (Pitch Synchronous-Overlap-and-Add) approach, described by Moulines and Charpentier (Moulines and Charpentier, 1990), is used. This approach is a time-domain technique; therefore, the modification of signal duration is performed directly in the time domain. The algorithm modifies the time position of speech frames according to some desired pitch contour. The consecutive steps of the PSOLA algorithm in the form of a pseudo-code are given below:

Step 1. Extracting pitch contour of speech sound

Step 2. Separating the frames

Step 3. Adding waves back

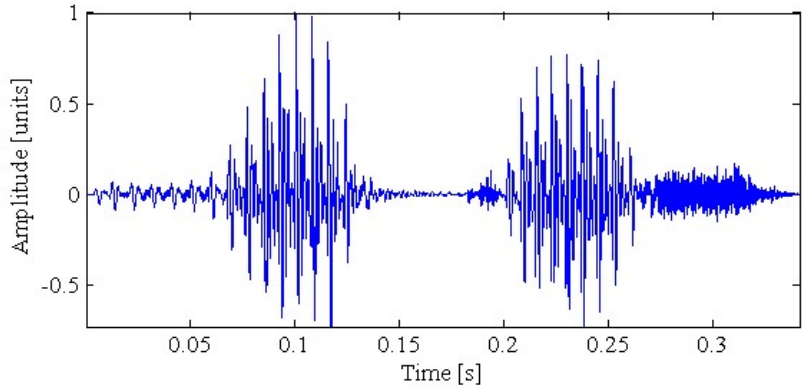
The applied frames are centered around the pitch marks and extended to the previous point and the next one. Modifications of duration are made in the following way:

- To increase the length of the signal, the frames are replicated.

- To decrease the length of the signal, the frames are discarded.

An example of the duration change is given in Figure 3.2.

a)




b)

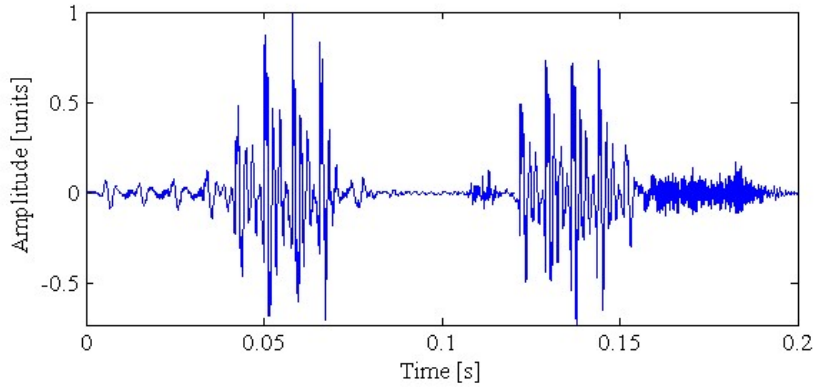



*Figure 3.2 The oscillogram of the Polish word "zakaz" (Eng. "prohibition"): a) before the vocal tract change, b) after decreasing the signal length by 40%*

**Applying modification to pitch**

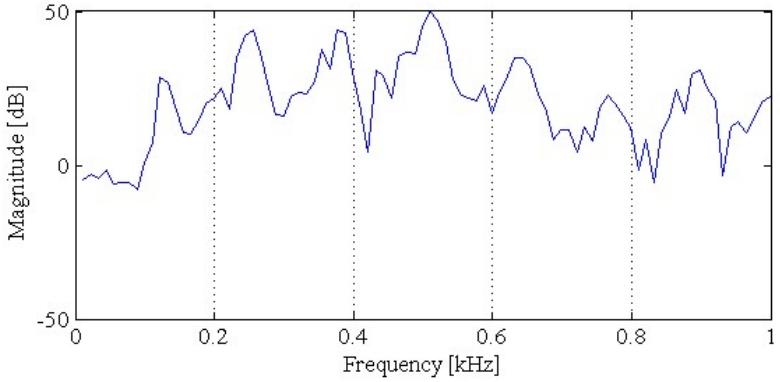
Modifications of pitch are achieved by varying fundamental frequency (F0) contour. Fundamental frequency was extracted using the autocorrelation method. The signal was divided into frames of the length of 0.05 s. Two manipulations of F0 contours are considered here. The first one is increasing and decreasing F0. The new pitch values are calculated by the following formula:

$$f_0^{(new)} = f_0 \frac{Mf_0^{(new)}}{Mf_0} \tag{3.1}$$

where symbol M denotes the median (the midpoint of the pitch array).

The pitch modification, as well as the process of the duration modification, is also implemented by employing the PSOLA approach, typically applied for such purposes. An example of the F0 increase is given in Figure 3.3.

a)




b)

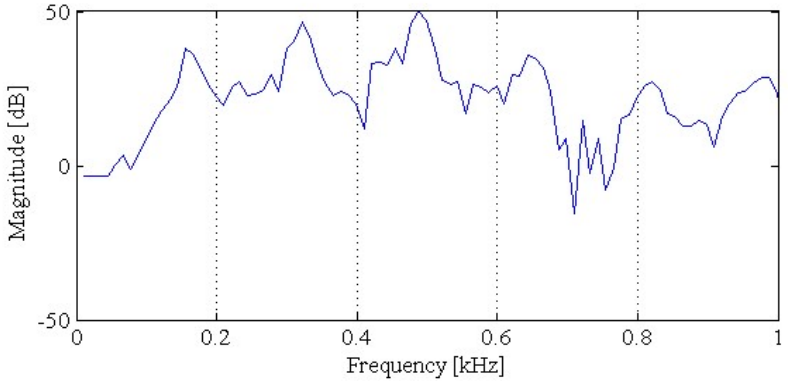



*Figure 3.3 Fast Fourier transform (FFT) spectrum of the Polish word "zakaz" (Eng. "prohibition"): a) before the F0 change, b) after increasing F0 by 40%*

The second manipulation, which is carried out, is called smoothing. This operation transforms the F0 contour into a continuous curve. In the experiments, the Praat built-in smoothing function is applied. Smoothing is achieved through changes in the underlying spectrum of the raw F0 contour that try to eliminate some of its higher frequency components (Arantes, 2015). An example of pitch smoothing is given in Figure 3.4. It should be noted that the modified signal retains the same duration as the original one.

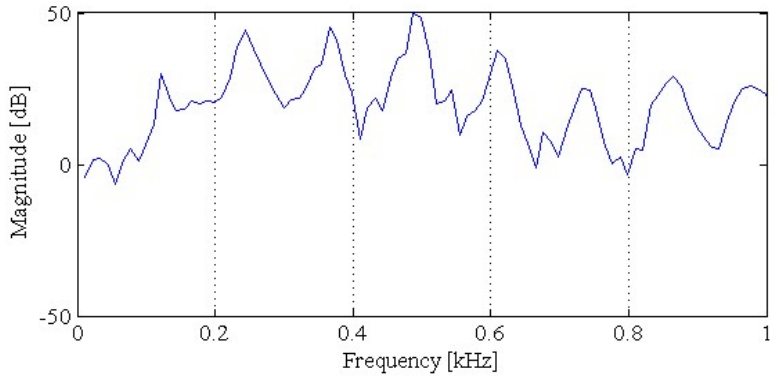



*Figure 3.4 The Fast Fourier transform (FFT) spectrum of the Polish word "zakaz" (Eng. "prohibition") after pitch smoothing*

**Modification of the vocal tract**

The goal of this modification is to scale the spectral envelope. The process of manipulation first begins with extracting the pitch contour of speech sound. Then the scaling coefficient is calculated by the following formula:

$$sc = \begin{cases} \frac{1}{1+\frac{x}{100}} & in \text{ the } case\ of \text{ the } increase \text{ of the } length\ of\ the\ vocal\ tract \\ 1 + \frac{x}{100} & in \text{ the } case\ of \text{ the } decrese \text{ of the } length\ of \text{ the } vocal\ tract \end{cases} \quad (3.2)$$

where $x$ is a decrease or increase in percentage.

For changing the vocal tract length, the four-step algorithm was applied. The algorithm was created based on the method given in the work by Darwin et al. (Darwin *et al.*, 2003). This method was used by Darwin et al. (Darwin *et al.*, 2003) to separate talkers of a different gender. The steps of the algorithm are given below:

Step 1. Multiplying pitch by sc

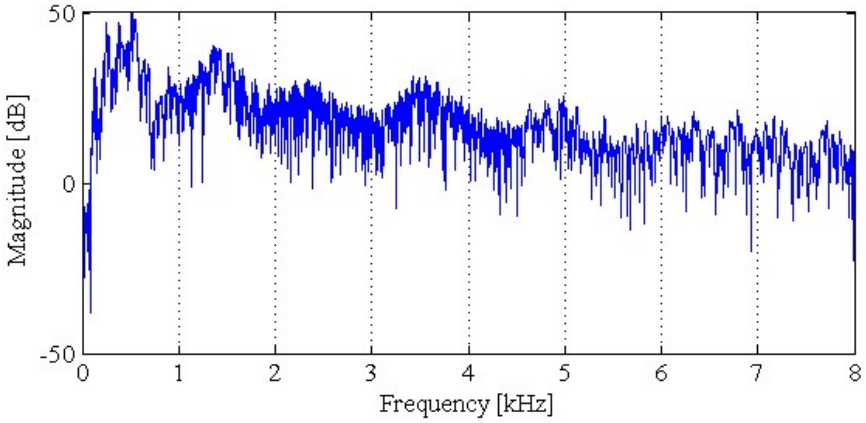
Step 2. Multiplying duration by 1/sc

Step 3. Resampling at the original sampling frequency multiplied by sc

Step 4. Playing the samples at the original sampling frequency.

To save the results of the vocal tract changing, the PSOLA approach was used. The graphical presentation of the vocal tract length changing is given in Figure 3.5.

a)




b)

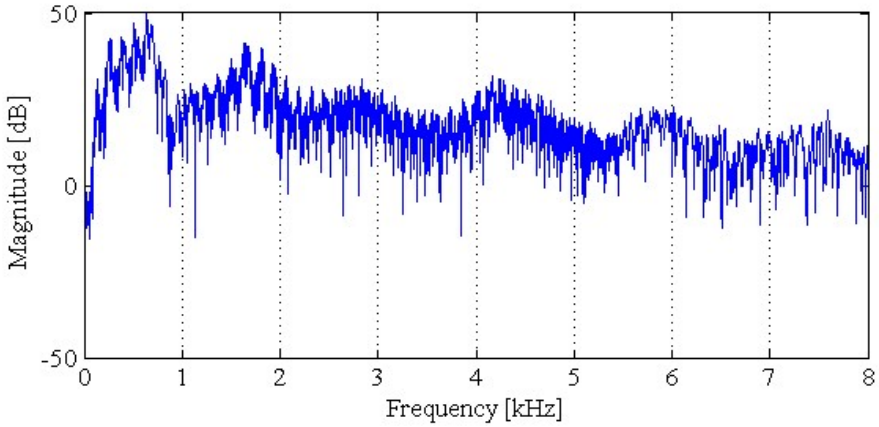



*Figure 3.5 The Fast Fourier transform (FFT) spectrum of the Polish word "zakaz" (Eng. "prohibition"): a) before the vocal tract change, b) after increasing the vocal tract length by 20%*

The modified signal, the spectrum of which is shown in Figure 3.4, retains the same duration and fundamental frequency as the original one.

**Manipulation of formants**

The speech signal can be considered as a combination of excitation signal (voice source) and vocal tract filter. The model describing this theory of speech production is the so-called source-filter model. The vocal tract filter is modeled based on formants (the peaks in the frequency spectrum). The excitation signal is constructed from the noise residual. In order to change formants, the LPC analysis, which decomposes speech sounds into the two above-mentioned parts, is performed.

There are several steps to be performed. First, the spectrum of the speech signal is filtered. The information below the lower edge f1 and above the upper edge f2 is removed (f1=0 and f2=5000 were assumed). Then the resampling of the speech signal from the real sampling frequency to frequency 2f2 is performed.

In the next step, the voiced parts of the speech signal are selected, and the LPC analysis is performed. The LPC analysis represents the spectrum with the chosen number of formants. The number of formants used in this research was equal to five. The LPC filter can be considered as an approximation to the vocal tract. The noise residual is obtained by filtering the speech signal through this LPC filter.

Then the formant frequencies of the speech signal are calculated. The process of calculation of formant frequencies is performed on the spectrum divided into short segments. Based on the Linear Prediction (LP) technique, each of these segments can be approximated as a linear combination of its $p$ previous samples:

$$\hat{x}(n) \approx a_1 x(n-1) - a_1 x(n-2) - \cdots - a_p x(n-p) \tag{3.3}$$

where $\hat{x}(n)$ is the predicted signal, $a = [1, a_1, \ldots, a_p]$ are a set of coefficients.

The formant frequencies are obtained as the roots of the polynomial given by Eq. (3.4). New formants are calculated by the following formulas:

$$F_1 = F_1 + \frac{x \cdot F_1}{100} \tag{3.4}$$

$$F_2 = F_2 + \frac{x \cdot F_2}{100} \tag{3.5}$$

where $x$ is a decrease in percentage.

The signal with modified formants is obtained by filtering the residual signal with formants given in Eqs. (3.4) - (3.5) through the filter with original LPC coefficients. In the last step of the modification procedure, reassembling of the speech signal to its initial frequency is performed.

## 3.2 Speech signal modeling techniques

This work focuses primarily on employing speech models to apply them in Lombard speech without changing parameters. This is because Lombard effect-related features should be incorporated into the speech signal. Previous investigations, in which the author of this dissertation participated, have shown that the Lombard speech model, based on dividing the speech signal into harmonics and modeling them as the output of a SISO (Single-Input and Single-Output) system

whose transfer function poles are multiple and inputs vary in time, retains the Lombard effect characteristics.

Representation of a speech signal by a sinusoid with time-varying amplitude and time-varying frequency is a prevalent method in speech modeling. A variety of techniques for synthesis in sinusoidal speech modeling have been proposed by researchers. The broad applicability of the sinusoidal approach is the main reason why this modeling technique is included in this dissertation. In speech recognition studies, the short-phase spectrum is still relatively infrequently taken into account. However, some scientists believe that the phase-based representation contributes to speech intelligibility just as much as the corresponding power spectrum. Moreover, Deng et al. pointed out that speech emotion recognition and speech enhancement areas may benefit from modifying the short-phase spectrum (Deng *et al.*, 2016). That is why, in this work, sinusoidal models without phase preserving and those with phase preserving are created to compare their efficiency in terms of speech quality measured in noisy conditions. An alternative to the sinusoidal paradigm is a source-filter model. The source-filter model is widely used in synthesizing human speech, as well as musical instrument sounds. This model is used for the research presented here because it is capable of synthesizing high-quality speech. Also, the source-filter model is implemented in the most popular vocoders belonging to the statistical parametric speech synthesis. However, it should be remembered that the aim of this dissertation is not to compare the vocoder implementation effectiveness but employ speech models for synthesizing the Lombard speech in the context of noisy conditions.

In the dissertation, the following models of speech: PSOLA, LPC, harmonic, source-filter, and sinusoidal, are to be investigated. Therefore, the description of each model is recalled in this Chapter.

### 3.2.1 Pitch-synchronous Overlap and Add (PSOLA)

PSOLA was designed to allow for modifying the prosody of natural speech with the ability to retain the speech naturalness. It was developed by Eric Moulines and Francis Charpentier (Moulines and Charpentier, 1990). Typical TT systems were then based on the concatenation of basic speech units, including diphones, demi-syllables, or non-uniform units – this approach was proposed by Sagisaka (Sagisaka, 1988). This approach requires an extensive database of acoustical units (Koszuta and Szklanny, 2017), and it is impossible in practice to provide such a database to synthesize the natural speech with natural prosody. One of the challenges is, therefore, to be able to change the pitch and duration of the speech.

The PSOLA synthesis flow consists of three steps:

- Analysis of the source speech waveform to provide the intermediate representation of the signal.
- Modifications made on this intermediate representation.
- Synthesis of the modified signal using the previously created modified intermediate representation.

This algorithm works in the following way:

- First, the speech waveform is divided into small overlapping segments.
- If the pitch needs to be changed, segments are either moved further apart to decrease the pitch or closer – to increase the pitch.
- If the duration must be changed, the segments are repeated multiple times to increase the duration, or some of them are removed to decrease the duration.

There are multiple variants of the PSOLA synthesis schema described by Moulines and Charpentier (Moulines and Charpentier, 1990).

### 3.2.2 Harmonic model

For harmonic modeling, a generator system proposed by Korvel and her colleagues (Korvel *et al.*, 2016, 2019) is used in this dissertation. The model is based on dividing the speech signal into harmonics and modeling them as the output of a SISO (Single-Input and Single-Output) system.

The impulse response $h_k(n)$ of the system is the 4th order quasipolynomial and is described by the following formula:

$$h_k(n) = e^{-\lambda_k n\Delta t} \sum_{m=1}^{4} a_{km} (n\Delta t)^{m-1} sin(2\pi k f_k n\Delta t + \varphi_{km}) \quad (3.6)$$

where $n$ is the discrete time, $\Delta t$ is the sampling period, $\lambda_k$ is the damping factor, $f_k$ is frequency, $a_{km}$ and $\varphi_{km}$ are amplitudes and phases, respectively ($m = 1, \dots, 4$; $k = 1, \dots, K$ ($K$ refers to the number of harmonics)).

The inputs of the $k^{\text{th}}$ harmonic can be described as follows:

$$u_k = [u_{k,1}, u_{k,2}, \dots, u_{k,M}] \quad (3.7)$$

where $u_{k,i}$ is the $i^{\text{th}}$ input of the $k^{\text{th}}$ harmonic and is calculated as the maximum amplitudes of the $i^{\text{th}}$ period of the $k^{\text{th}}$ harmonic. The detailed procedure for determining inputs and distances between them is presented in the paper by Pyž et al. (Pyž *et al.*, 2014).

### 3.2.3 Source-Filter Model

Based on the source-filter theory, the speech signal is produced by an excitation, which is then filtered by a vocal tract shape. The vocal tract filter can be described as a linear time-invariant system. The mathematical expression of the speech signal model, denoted by $y(t)$, which is an output signal of such a system, is as follows:

$$y(t) = h(t) * x(t) \tag{3.8}$$

where $*$ denotes the convolution operation,

$$x(t) * h(t) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \tag{3.9}$$

and $h(t)$ is the impulse response of the system, and $x(t)$ is the input signal.

In this research, to achieve a high-quality speech model, two models based on different architectures are constructed. In both of them, the input signal is a pulse train with the fundamental period.

**Linear Predictive Coding (LPC)**

LPC is a well-known technique of speech analysis and synthesis. LPC is based on a source-filter model, where the source is basically a spectrally flat signal: impulse train - used to model the pitched sounds – or random white noise to model plosives or fricatives (as a basic version). Most variants of LPC use, however, more complex excitation signals. The resonant filter models the vocal tract. In other words, LPC flow models the glottis producing the buzz, characterized by intensity and pitch frequency, and the throat and mouth forming the tube, characterized by resonances, that are producing formants.

The filter used in the LPC synthesis and the buzz signal is unknown and must be estimated using the speech signal. That is why the important part of LPC is analysis. In its analytical part, LPC estimates the formants, removes their effects from the speech signal, and estimates the frequency and intensity of the remaining buzz. The remaining signal after removing the formants is called the residue. The filter coefficients are calculated using the linear combination of the previous outputs – in other words, they are "predicted," hence the name of the algorithm is Linear Predictive Coding.

The reverse process – synthesis – is performed using the residue, buzz parameters, and formants, creating a filter. The source is passed through the filter, and the speech signal is produced. This process is performed on small chunks of the speech signal to allow for generating longer speech signals.

**Source-filter model with aperiodicity parameter**

In the source-filter model, the excitation depends on a fundamental frequency); therefore, first, the contour is estimated. For this purpose, a method based on both time interval and frequency cues is used (Kawahara *et al.*, 2005). This method provides fundamental frequency and periodicity information within each frequency band. The aperiodicity information is estimated from the residuals between harmonic components and is used for synthesizing both periodic and aperiodic signals. Although in the source-filter theory, a source signal and a vocal-tract filter are separated, in real conditions, there is an interaction between them. Therefore, a parameter is included in the spectral envelope estimation algorithm. The basic principles of this algorithm can be found in the paper by Kawahara (Kawahara, 2006). The algorithm extracts a smoothed time-frequency representation. The reconstructed spectrogram is commonly known as a STRAIGHT spectrogram (Kawahara, 2006).

In a STRAIGHT implementation, the STRAIGHT spectrogram, aperiodicity map, and fundamental frequency F0 with voicing information are used in the process of speech synthesis. These parameters allow for independent manipulation without introducing any inconsistency between them. This means that it is possible to set the pitch, timbre, duration, or any other parameter of speech flexibly. An overlap-add synthesis using minimum-phase impulse response with group delay manipulations is used for this purpose.

The schema of the STRAIGHT vocoder is presented in Figure 3.6.

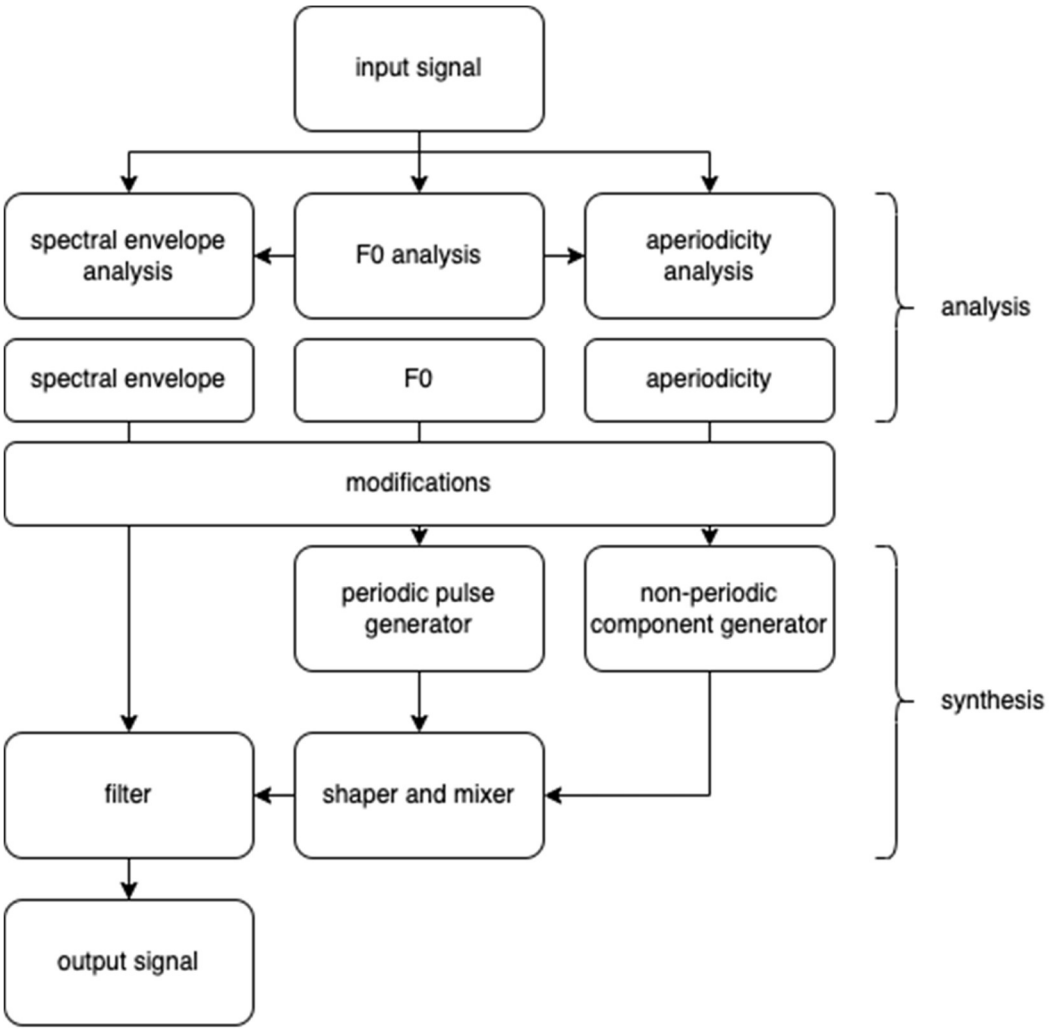



*Figure 3.6 Schema of the STRAIGHT vocoder*

**Source-filter model with waveform-based parameter**

It is well-known that the human voice is not perfectly periodic. That is why in speech synthesis, the mixed excitation signal containing an aperiodic signal should be applied. According to Morise, when a periodic signal is calculated as the minimum-phase response, the model cannot represent the phase of the input voice as the vocal tract response generally includes not only a minimum-phase reaction but also a maximum-phase response (Morise, 2012; Morise *et al.*, 2016). The author pointed out that to accurately synthesize a voice, it is essential to extract the phase of the input signal.

In this model, an instant of aperiodicity, a waveform-based parameter, is used. The model is realized using a high-quality speech analysis, modification, and synthesis system developed by Morise et al. (Morise *et al.*, 2016). It consists of three analysis algorithms for obtaining speech parameters and one synthesis algorithm that takes these parameters as input. In the process of

analysis, first of all, the parameter and spectral envelope are estimated. As in the case of the source-filter model described above, the information is also used in the spectral envelope estimation process. The fundamental frequency, spectral envelope information, as well as signal waveform are used for the estimation of the excitation signal. During the modeling process, these estimations are incorporated. The details of the algorithm implemented are presented in earlier works by these authors (Morise, 2012, 2015; Morise *et al.*, 2009).

The algorithm developed by Morise et al. (Morise *et al.*, 2016) is further referenced as the WORLD vocoder.

This vocoder is used in the experiments both to gather the information about the speech fundamental frequency and to resynthesize the speech signal using the changed F0 value. One of the important factors impacting the resynthesized speech quality is the algorithm of the fundamental frequency refinement called Stonemask (Morise, 2016). As a fundamental frequency estimator, Harvest is used (Morise, 2017) and as an aperiodicity estimator – D4C, developed by Morise (Morise, 2016).

In this dissertation, the WORLD vocoder is used in the experiments since it allows for real-time speech processing (Morise *et al.*, 2016).

Contrary to STRAIGHT, which calculates the vocal cord vibration independently from the periodic and aperiodic responses, in WORLD, the vocal cord vibration is calculated using the convolution of minimum phase response and the extracted excitation signal.

The most important feature of the WORLD vocoder is its low computational cost. It is because it has fewer convolutions than STRAIGHT and, therefore, can work in real-time.

Figure 3.7 presents how the speech signal is analyzed and resynthesized using the WORLD vocoder in the main experiments of this dissertation. For this purpose, Python code developed by Jeremy Hsu (Hsu, 2021) is utilized.

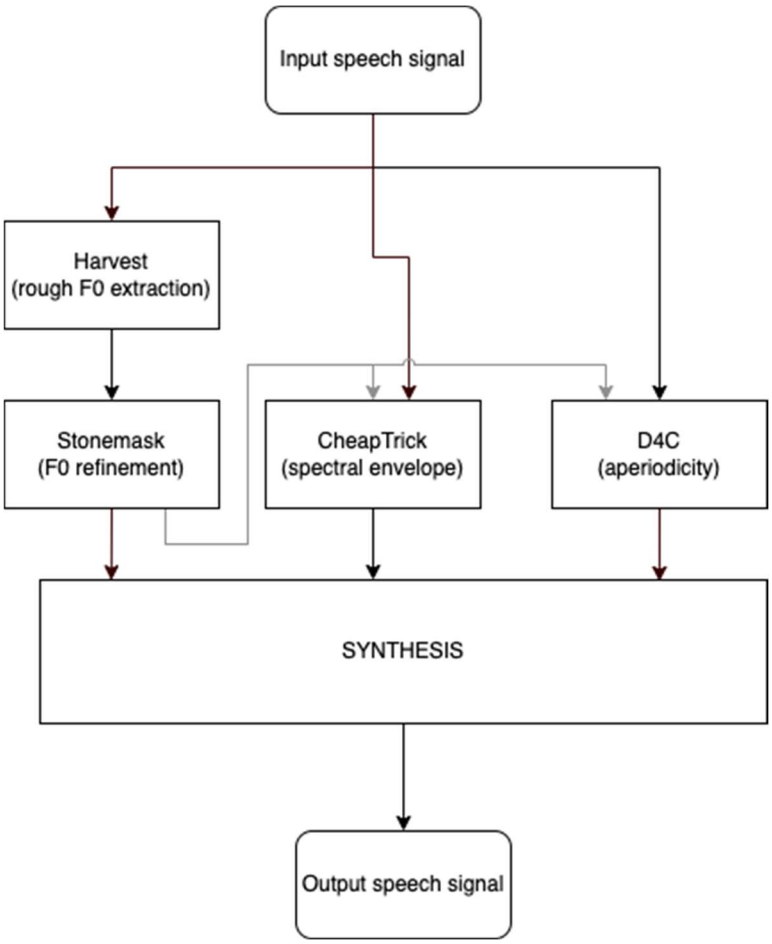



*Figure 3.7 Speech signal processing method using the WORLD vocoder, implemented in this dissertation*

### 3.2.4 Sinusoidal model

According to the sinusoidal speech modeling technique, the signal is represented as a sum of sinusoids whose frequency and amplitude vary in time. In this research, parameters of sinusoids are determined by tracking the Fast Fourier Transform (FFT) spectral peaks, as per the example given in Ellis (Ellis, 2003).

**Sinusoidal model without phase preserving**

The construction of the model begins with the construction of the sinusoidal representation of the speech signal. For this purpose, the Short-time Fourier Transform (STFT) spectrogram, which is a visual representation of the signal spectrum that varies with time, is used.

Based on the spectrogram, a speech analysis is performed, which determines the stationary and deterministic parts of the speech signal. For this purpose, frequencies and amplitudes corresponding to local peaks in the spectrum are detected. The other task is to determine which peaks belong to the spoken signal. To achieve that, a list of detected peaks is fed into the tracking

procedure. According to this procedure, for each frequency $\omega_i^k$ in frame $k$ we are looking for the frequency $\omega_j^{k+1}$ in frame $k$+1 is sought, which is closest to such a frequency and whose absolute distance is less than the threshold Δ, i.e.:

$$\left|\omega_i^k - \omega_j^{k+1}\right| < \left|\omega_i^k - \omega_p^{k+1}\right| < \Delta \tag{3.10}$$

where the $|.|$ symbol refers to absolute value or magnitude, $i = 1, \dots, L_k$ ($L_k$ – the total number of peaks in frame $k$), $j = 1, \dots, L_{k+1}$, ($L_{k+1}$ – the total number of peaks in frame $k + 1$), and $(p = 1, \dots, L_{k+1}) \cap (p \neq j)$.

If the match between frequencies is not found, they are matched to themselves, and their magnitudes are set to zero. As a result, an interpolated peak magnitude for each track point is obtained.

In the last step, speech signal resynthesis is performed. For reconstruction, a sine wave oscillator bank developed by Ellis is used (Ellis, 2003).

**Sinusoidal model with phase preserving**

Most speech processing applications are based on the short-time spectrum, while relatively little attention is paid to the short-range phase spectrum. According to Abe and his colleagues, the instantaneous frequency (IF) spectrogram more clearly shows the harmonic structure of quasi-periodic signals such as speech than STFT spectrograms (Abe *et al.*, 1997). The advantages of including phase-related information in the speech vocoder are listed in these works (Ellis and Weiss, 2006; Laroche and Dolson, 1999).

In this model, instead of the linear spectrogram, which discards phase information, the IF spectrogram is used. The harmonic frequencies based on IF of a speech signal are obtained by the technique proposed in the work by Abe and his colleagues (Abe *et al.*, 1997).

Resynthesizing consists of reading the series of frequency, magnitude, and phase samples for a particular track. For this purpose, the Matlab code developed by Ellis (Ellis, 2003) is utilized.

## 3.3 Machine learning approach to speech modification

Speech modification using machine learning methods is an efficient method of adaptation of the speech features to different conditions (López *et al.*, 2017; Seshadri *et al.*, 2019). While using the previously mentioned speech modification methods or any other vocoder, the speech features can be dynamically updated to reflect, for instance, a given style of speech such as Lombard (Bollepalli *et al.*, 2019).

There are different approaches to this topic, but most of them are applied to text-to-speech (TTS) systems (Bollepalli *et al.*, 2019; Raitio *et al.*, 2011). This is because the speech (text) features should be known before they are transformed.

In this work, none of the mentioned above approaches are used since this dissertation aims to manipulate speech without knowing its features based on text. Thus, these methods are only shortly described to draw the complete picture of processes allowing to adapt the normal speaking style to the Lombard one. Although they are perfect in TTS systems, when the phonemic contents and the text features are known, they cannot easily be used in real-time systems if only the recording of the speech or live speech is available.

### 3.3.1 Auxiliary features

It might be assumed that any deep acoustic model is used to transform text features into speech output or – to be more precise – into acoustic features of speech (either neutral (normal) or Lombard). The internal values of the model (for instance, weights of the deep neural network) are trained on the normal speech, and there is a need to change these values. However, the auxiliary features method takes another approach – the input vector of text features is changed so that the new – auxiliary – features are appended. Then the internal model does not have to be changed at all. Usually, for Lombard speech adaptation, a one-hot vector indicating the speech style (Lombard or neutral) is added to the text features (Bollepalli *et al.*, 2019).

### 3.3.2 Learning hidden unit contribution (LHUC)

The Learning hidden unit contribution (LHUC) method takes the same assumption as auxiliary features but does not change the input text features vector. Assuming that the neural network is trained on data derived from different speakers representing different styles, the hidden network represents the acoustic model of various speech styles.

Taking the initially learnt parameters, a new set of parameters might be discovered that represent the Lombard speech style. Then the initial network parameters are rescaled using the Lombard speech representation (Bollepalli *et al.*, 2019).

### 3.3.3 Fine-tuning

A fine-tuning method is a transfer learning approach (Pan and Yang, 2010). Assuming that there is a neural network representing normal speaking style, there is a possibility of re-training using the Lombard speech dataset.

Typically the learning process starts from an average voice model – represented by the neural network, pre-trained using the large dataset of speakers. Later the model is upgraded by training it

using the smaller, specific dataset. In other words, the existing, pre-trained network learns most of the relations between text and acoustic features, and there is only a need to update the network to reflect the specific speaker (Bollepalli *et al.*, 2019). Fine-tuning has been presented in multiple works (Cooper and Hirschberg, 2018; Takaki et al., 2016), and its overall efficiency was demonstrated (Takaki *et al.*, 2016).

## 3.4 Evaluation methods

### 3.4.1 ITU standard-based evaluation

From the point of view of speech quality, the best measure of quality is the subjective measurement of intelligibility. ITU (International Telecommunication Union) defines standards for subjective measurements - using listeners' experts. Two of them are very important in the above context, i.e., P.800 / P.830 standards (ITU-T Recommendation P.800, 1996; ITU-T Recommendation P.800.1, 2006). The results of this type of measurement are presented as MOS (Mean Opinion Score) (ITU-T Recommendation P.800.1, 2006). Such evaluation should be performed as listening tests on a group of subjects. Despite the standardization of this type of tests (requirements regarding the acoustics of the listening room, admissible interference value, monitoring system, way of testing, reliability of testers, etc., included in the standards, among others: ITU-R BS.1116 (2016) and ITU-R BS. 1284 (2003), they are usually subject to errors since each listener may have different auditory experiences. However, it should be remembered that speech intelligibility is, by definition, a subjective indicator, which is why this type of research is conducted as the final verification of the results obtained.

However, the quality of the speech signal can also be measured objectively (ITU-T Recommendation P.563, 2004). Objective indicators regarding the speech signal (e.g., clarity) are most often used in the measurement of the quality of telecommunications channels. It is important primarily because of the need to ensure the proper quality of services.

There are many factors that influence the speech signal quality, including:

- a narrow transmission band or coding using a low transmission rate,
- compression and coding algorithms,
- background noise,
- delay in packets in digital transmission,
- quality of transmission devices (e.g., mobile phones).

It should be noted that the quality of speech in telecommunications can be assessed by (Beerends *et al.*, 2002; Beerends *et al.*, 2013; ITU-T Recommendation P.563, 2004):

- double-ended measurement – this type of measurement refers to comparing the reference signal and tested signal (i.e., before and after the channel),
- single-ended measurement – this type of quality measurement considers and estimates speech quality perceptual aspects without knowledge of the reference signal.

**PESQ MOS-based measurement method**

Historically, P.861, known as PSQM (Perceptual Speech Quality Measure) (Beerends *et al*., 2022), was the first standard for measuring speech signal quality. At that time, it did not consider many aspects of modern digital transmission channels, e.g., loss of packets in VoIP (Voice over Internet Protocol), background noise or interferences, overhead delays, variable delay, etc., that can affect the measurement. Conversely, P.862 standard (PESQ) (ITU-T Recommendation P.862, 2001) took these issues into account. What is also important is that when comparing objective and subjective test results, the correlation coefficient between PSQM and subjective results returned the value of 0.26, while in the case of the PESQ-based analysis reached 0.93. This is because, in PESQ, the original and degraded signals are mapped onto an internal representation using a perceptual model. Beerends *et al*. (Beerends *et al.*, 2013) showed that a cognitive model uses the difference in this representation to predict the perceived speech quality of the degraded signal.

The PESQ algorithm is realized through several steps, which are listed below:

- the original and degraded signals should be aligned to the same power level,
- signal filtering modeling; this concerns telephone devices and the telecommunication networks,
- delay compensation, i.e., time alignment as transmission systems may introduce a significant delay to the signal,
- the auditory perceptual transformation system processes both the reference and the degraded signals to simulate the characteristics of human hearing,
- the calculated disturbance parameters are converted into the PESQ scores ranging from -1 to 4.5. This adheres to the MOS-LQO (Listening Quality Objective) scale (ITU-T Recommendation P.862.1, 2003), i.e., values from 1 to 5 (where scores are interpreted as follows: “1” means bad quality of speech and “5” refers to excellent quality).

In conclusion, based on the above considerations, Perceptual Evaluation of Speech Quality, PESQ is a measure of signal quality in the telecommunication channel (ITU-T Recommendation P.862, 2001). However, PESQ can also be employed for objective speech quality tests in the presence of environmental or ambient noise.

In this dissertation, the PESQ algorithm available on the ITU websites is implemented. This enables to carry out the PESQ measurement by comparing two signals, i.e., the original, treated as

a reference signal, and a test signal that contains interference. For the purpose of experiments, speech signals mixed with pink noise or a babble speech signal mimicking the buzz of human voices were employed.

**ITU-T Recommendation P.563**

In single-sided measurements, the MOS value is estimated exclusively on the basis of the interference signal. In the case of the P.563 standard, the use of a real expert listening to the conversation on the test device should be simulated. This device can be any receiver, e.g., a mobile phone. Since, in this case, the degraded signal is not compared to the original signal, the speech quality indicator depends on the listening device. It is, therefore, an important element of the P.563 standard (ITU-T Recommendation P.563, 2004).

Each signal subjected to MOS measurement using P.563 must be pre-processed by using the model of the listening device. Next, a speech detector (VAD - Voice Activity Detector) is used to mark the speech-related signal fragments. In the next stage, the speech signal is subjected to a series of analyses and assigned to a given class of disturbances. Parameterization of the signal in P.563 can be divided into three basic function blocks (ITU-T Recommendation P.563, 2004):

- analysis of the vocal tract and speech abnormality; in this case, it is possible to distinguish the indication of unnaturalness separately for female and male voices and the so-called the "robot" effect,
- additional noise analysis; in this case, it is important to detect constant background noise and noise associated with the signal envelope,
- interruptions, mute, and time cuts.

The test signal must also meet the requirements specified in the standard so that there is the possibility of detecting speech quality using the P.563 algorithm, including:

- the sampling frequency must be greater than or equal to 8 kHz,
- the digital signal resolution must be 16-bit,
- the signal cannot be longer than 20 seconds, and the speech in the signal cannot be shorter than 3 seconds.

Using P.563 as a speech quality metric often enables quick and reliable system adaptation, taking the channel quality into consideration. As an output, P.563 measurements return Mean Opinion Score - Objective Listening Quality (MOS-LQO), which shows quite a high correlation with the MOS-LQS values returned from the subjective tests (ITU-T Recommendation P.563, 2004).

Some researchers show a high correlation between the MOS-LQS and MOS-LQO obtained with the P.563 algorithm (Dubey and Kumar, 2015; Kraljevski *et al.*, 2010), even though different

speech characteristics can be evaluated; for instance, speech naturalness or intelligibility. Correlation between P.563 measurement and subjective tests also depends on the speech sampling frequency - comparisons show that 16 kHz sampling frequency gives a better correlation in terms of naturalness and intelligibility (Kraljevski *et al.*, 2010).

The basic block scheme of the P.563 algorithm is shown in Figure 3.8.

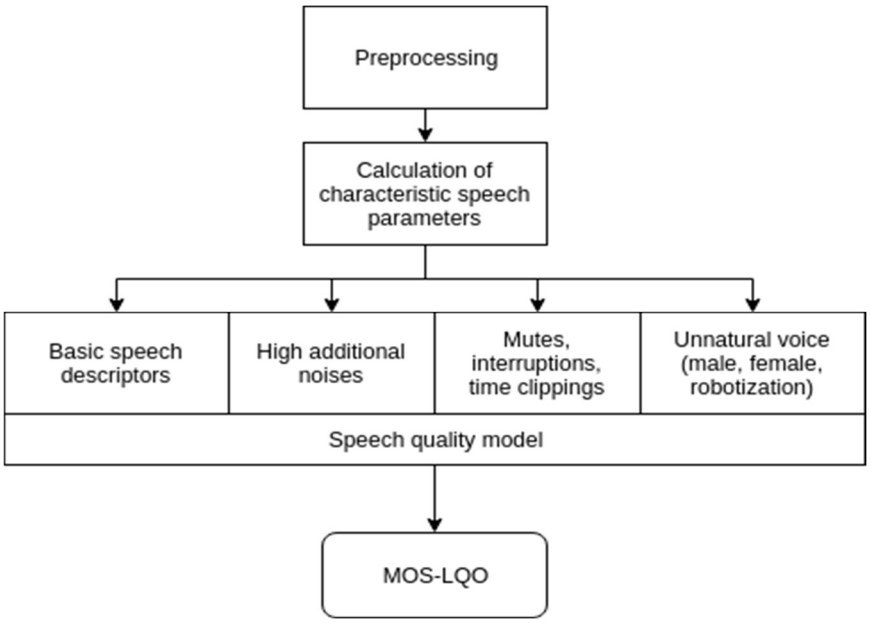



*Figure 3.8 The basic block scheme of the P.563 algorithm (ITU-T Recommendation P.563, 2004)*

**Speech quality measure based on acoustic parameters**

As already said, the ITU standards concern speech quality measurement in the telecommunication channel, so they are not best suited to detecting and testing the Lombard effect in a speech-in ambient noise conditions. Therefore, a novel method was proposed for that purpose, based on a set of parameters presented in an earlier co-authored work (Korvel *et al.*, 2020). Moreover, a measure of speech model (discussed in Chapter 3.1) quality, depending on these parameters, was introduced. It was worth mentioning that such a quality indicator is particularly well-fitted for synthesized speech models. The list of the signal descriptors employed for that purpose is given in Table 3.1.

*Table 3.1 The acoustic parameters for evaluation of the Lombard effect in models*

| **The time-domain parameters** | |
|---|---|
| 1 | Temporal Centroid |
| 2 | Zero Crossing Rate |
| 3 | Root Mean Square (RMS) energy |
| 4-6 | The number of samples exceeding levels RMS, 2×RMS, 3×RMS |
| 7-12 | The mean and variance of samples exceeding levels RMS, 2×RMS, 3×RMS averaged for 10 sub-segments |
| 13 | Peak to RMS |
| 14-17 | The number of the signal crossings in relation to zero, RMS, 2×RMS, 3×RMS |
| 18-25 | The mean and variance of signal crossings in relation to zero, RMS, 2×RMS, 3×RMS averaged for 10 sub-segments |
| **The frequency-domain parameters** | |
| 26-30 | The first five formants |
| 31 | Audio Spectral Centroid (MPEG 7 standard) |
| 32 | Audio Spectral Spread (MPEG 7 standard) |
| 33 | Audio Spectral Skewness (MPEG 7 standard) |
| 34 | Audio Spectral Kurtosis (MPEG 7 standard) |
| 35 | Spectral Entropy (MPEG 7 standard) |
| 36 | Spectral Roll-Off (MPEG 7 standard) |
| 37 | Spectral Brightness (MPEG 7 standard) |
| 38-66 | Audio Spectrum Envelope calculated on 29 sub-bands (MPEG 7 standard) |
| 67 | Mean Audio Spectrum Envelope (MPEG 7 standard) |
| 68-85 | Spectral Flatness Measure calculated on 18 sub-bands (MPEG 7 standard) |
| 86 | Mean Spectral Flatness Measure (MPEG 7 standard) |
| 87-106 | Mel-Frequency Cepstral Coefficients |

The parameters include time- and frequency-domain features (see Table 3.1). The frequency-domain parameters are calculated from the FFT spectrum. The speech signal is divided into short-time segments with a 50% overlap, and Hamming window is applied to each segment before the derivation of the parameters.

The proposed measure of speech quality is based on normalized distances between parameters. The distances are described by the following formula:

$$Dist = \sum_{i=1}^{N} \frac{|SD_i(Lomb) - SD_i(model)|}{Max\_par_i} \tag{3.11}$$

where $N$ is the number (i.e., $N = 106$) of parameters, $SD_i(Lomb)$ is the standard deviation of $i^{th}$ Lombard speech parameter vector calculated on short-time segments and $SD_i(model)$, the standard deviation of the $i^{th}$ parameter vector of the model derived from short-time segments is calculated as follows:

$$SD_i = \sqrt{\frac{\sum_{j=1}^{M}(r_{ij}-\overline{r_i})^2}{M-1}} \tag{3.12}$$

where $M$ is the number of short-time segments, $r_{ij}$ – the $j^{th}$ value of the $i^{th}$ parameter vector, $\overline{r_i}$ –the mean value of the $i^{th}$ parameter vector.

Parameter $Max\_par_i$ is calculated as the maximum value of the $i^{th}$ parameter vector of natural speech, i.e.:

$$Max\_par_i = max\ \{r_{ij}(Lomb)\} \tag{3.13}$$

In the last step of the measure construction, the distances (see Eq. 3.8) are normalized to the interval [1, 5], corresponding to the MOS-LQS scale.

### 3.4.2 Subjective tests

When it comes to speech models, a subjective test is an essential element of the evaluation process that allows assessing the quality of the obtained sounds. Therefore, the subjective evaluation of speech models was also performed. This evaluation was based on a modified MUSHRA listening test. The modification applied will be explained later on.

MUSHRA stands for MUlti Stimulus test with Hidden Reference and Anchor. It is a test that allows a subjective comparison of multiple audio signals, suitable for intermediate audio quality (ITU-T Recommendation BS.1534-1, 2003). MUSHRA is described in ITU recommendation BS.1534-1 and updated in BS.1534-2.

There are some requirements that describe the MUSHRA test, for instance:

- The sequence should not exceed 20 s to avoid fatiguing listeners and to reduce the total duration of the listening test.
- In total, a session should not last for more than 20 minutes to avoid fatigue in judgments.
- The set of signals should contain one reference signal (full quality) and one low-pass filtered signal version (the so-called anchor, typically with 3.5 kHz bandwidth); additional anchors might be used optionally.

Despite the MUSHRA usability, one should be aware of potential biases that may occur when preparing test signals and constructing the whole set to be evaluated (Zielinski, 2016). In the designed experiments, Zieliński showed systematic discrepancies in the results of the MUSHRA

test (Zielinski, 2016). They refer to stimulus spacing bias, centering bias, range equalizing bias, contraction bias, and bias due to the nonlinear properties of an assessment scale. The possible biases that may occur in the tests performed will be discussed in the Conclusion Chapter.

The MUSHRA test used in this dissertation was created by the author using the web interface and the Audiolabs' MUSHRA application (webMUSHRA, 2019). It has been installed on a web server and configured using the following assumption: every page in the MUSHRA test contains a single sentence of a single person with different types of modifications with the same level and type of noise.

The test is available online. It was, however, modified to adapt to the quality of the presented signals. Test users in the pre-test session reported that the clean (reference) signal and the low-pass filtered anchors disturbed the overall listening experience during the test, thus not allowing for the correct quality assignment. Therefore, the reference signal and the anchor were removed. That is why, in this work, the test is referred to as a "modified MUSHRA test."

For the statistical analysis of data obtained with the MUSHRA method, the ANOVA test, which is supported by the recommendation, was used (ITU-T Recommendation BS.1534-1, 2003).

# 4 Experiments

## 4.1 Preliminary experiments

### 4.1.1 Motivation

Preliminary experiments were performed to define how the speech signal can effectively be modified to improve its quality in noise conditions. There are different methods that may be applied to this end – for instance, vocoders (Kawahara, 2006; Morise *et al.*, 2016) or other types of signal transformation. Some experiments were performed to be able to identify the best method, used later in the proposed adaptive system. Preliminary experiments are, in fact, a simplified approach, allowing to define the best strategy for an adaptive system.

For that purpose, speech recordings with the presence of noise were used. In the preliminary experiments, the recordings from an audio-visual corpus for multimodal automatic speech recognition were used (Czyzewski *et al.*, 2017). However, the sound files do not contain noise itself – according to the description of the recording method. Therefore, the recordings were mixed with the pink noise of a different signal-to-noise ratio (SNR). For the consecutive samples SNR was set respectively to -10 dB, -5 dB, 0 dB, 5 dB and 10 dB.

In order to improve objective quality indicators of speech utterances in noise conditions, various signal modifications were considered. For the purpose of preliminary experiments, Praat software scripts created by Corretge (Corretge, 2012), their modification introduced by the author, as well as new scripts designed within this work were used. All measurements and manipulations were made with the Praat Software version 6.0.39 (Boersma and Weenink, 2018). A detailed description of signal modifications is given in this Chapter. The focus was on frequency-domain filtering, manipulation of duration, applying modification to pitch, modification of the vocal tract, and manipulation of formants.

### 4.1.2 Improving PESQ MOS measures – experiment 1

**Mixing the recording with pink noise and babble speech**

As already mentioned, the processing of the recorded utterances encompasses two types of distortions that were mixed with the recording:

- pink noise,

- babble speech, which is a typical noise for places such as cafeterias.

Both types of distortions were mixed with the original signal with the same signal-to-noise ratios (SNR) as when comparing recordings without and with Lombard speech, i.e., -10, -5, 0, 5, and 10 dB.

**PESQ estimation**

The following charts present the results of MOS estimation using the PESQ algorithm. Every bar represents one proposed modification of the input signal. The resulting estimated MOS values should be compared with the ones obtained by measuring the speech quality of the clean, non-Lombard speech signal. Therefore every chart presents the obtained PESQ MOS values for different distortion types and different levels of SNR. The modification types are described using the following denotations (see Table 4.1) contained in the charts.

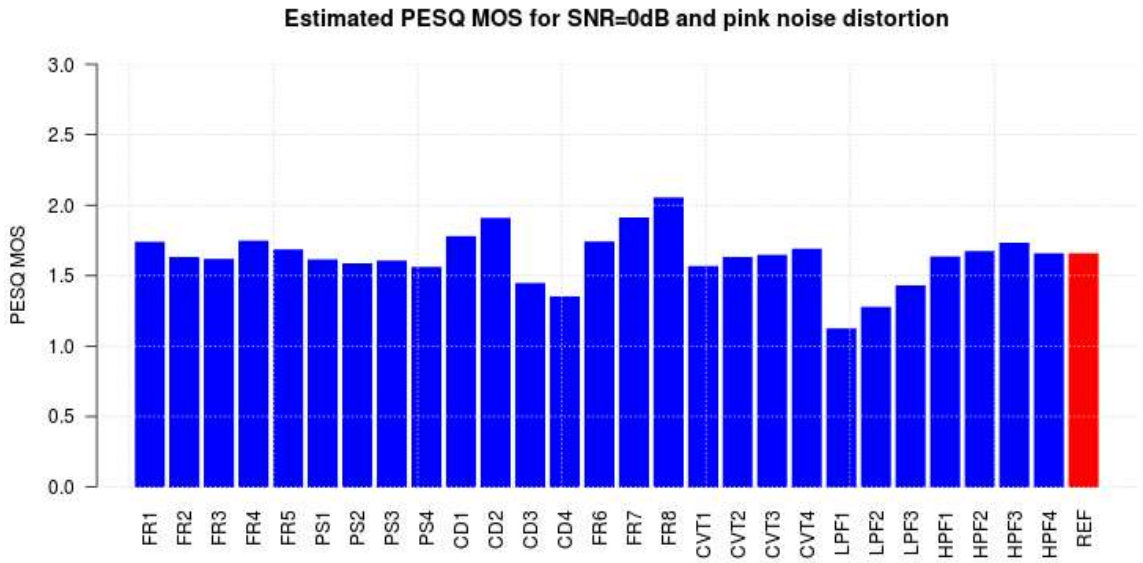
*Table 4.1 Denotation concerning modification types of the input signal (REF - reference signal, i.e., original speech recorded in silent conditions)*

| Increase of formants F1 and F2 in [%] | Pitch smoothing in [%] | Increase of duration in [%] | Vocal tract moving to the left or the right, given in [%] | Low pass filter with a cutoff frequency in [kHz] | High pass filter with a cutoff frequency in [Hz] |
|---|---|---|---|---|---|
| FR1 – F1 and F2 raised respectively by 10% and 8% | PS1 – 100% | CD1 – 20% | CVT1 – moving 20% to the left | LPF1 – 6 kHz | HPF1 – 120 Hz |
| FR2 – F1 and F2 raised respectively by 3% and 2% | PS2 – 25% | CD2 – 40% | CVT2 – moving 10% to the left | LPF2 – 7 kHz | HPF2 – 180 Hz |
| FR3 – F1 and F2 raised respectively by 4% and 2% | PS3 – 50% | CD3 – 20% | CVT3 – moving 10% to the right | LPF3 – 8 kHz | HPF3 – 240 Hz |
| FR4 – F1 and F2 raised respectively by 5% and 4% | PS4 – 75% | CD4 – 40% | CVT4 – moving 20% to the right | | HPF4 – 60 Hz |
| FR5 – F1 and F2 raised respectively by 8% and 6% | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| FR6 – F1 and F2 raised by 10% | | | | | |
| FR7 – F1 and F2 raised by 20% | | | | | |
| FR8 – F1 and F2 raised by 30% | | | | | |

Examples of the analyses are shown below. They compare speech signal modifications with the reference signal (original speech recorded in silent conditions). Figure 4.1 shows examples of estimated PESQ MOS values for SNR=0 pink noise (a) and babble speech distortions (b). Some modifications seem to be more effective than others. For instance, raising the formants and changing duration positively impact the estimated MOS. Other such analyses (SNR: -10 dB up to +10 dB; pink noise/babble speech) show that regardless of the noise applied and SNR, this tendency stays valid for all other cases. Below, charts from Figure 4.2 shows only four modifications – CD1, CD2, FR7, and FR8 for additional comparison. In all cases of analyses, PESQ MOS values are higher for pink noise disturbance than for babble speech; however, the differences between them are relatively small. That is why Figure 4.3 compares the results of processing the most effective changes (i.e., raising formants) as a function of SNR. In all cases, such a change applied improves MOS for most SNRs. Overall, changes in MOS are present for both positive and negative SNRs.

a)



b)

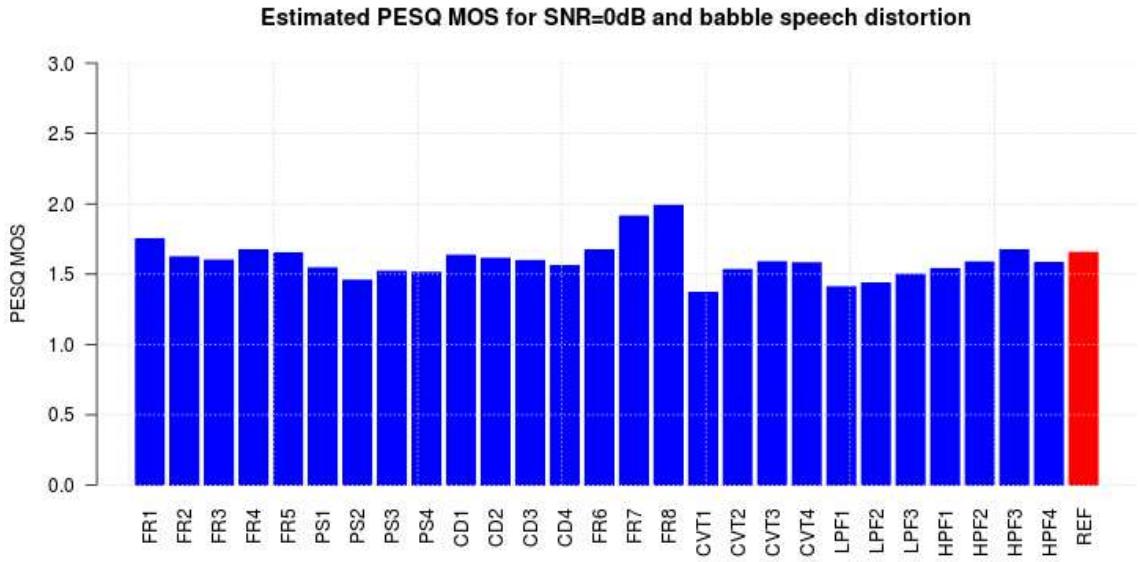
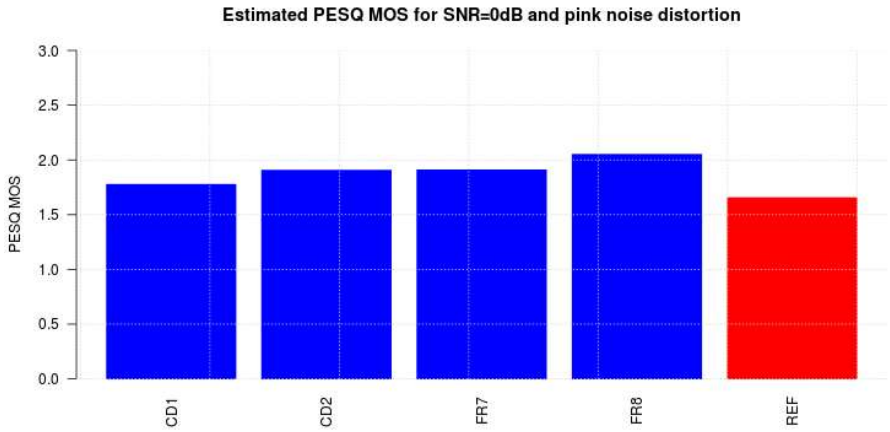



*Figure 4.1 Estimated PESQ MOS values for SNR=0, pink noise (a) and babble speech (b) distortions*

a)



b)

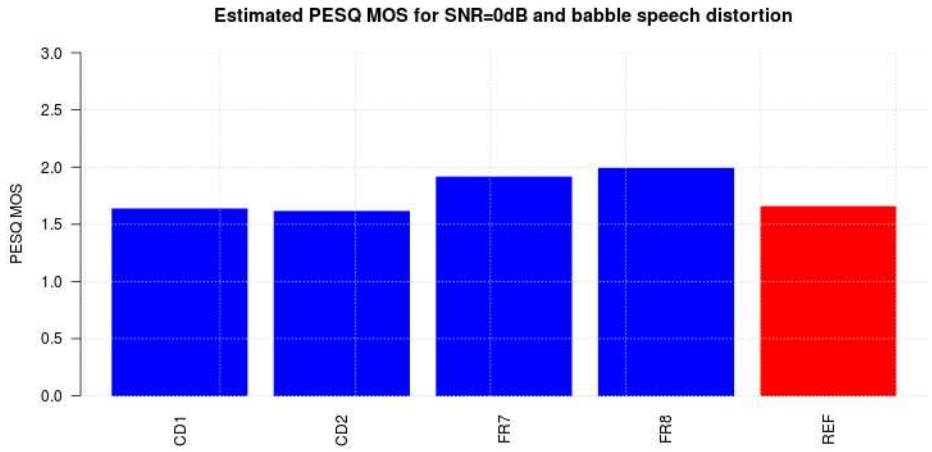
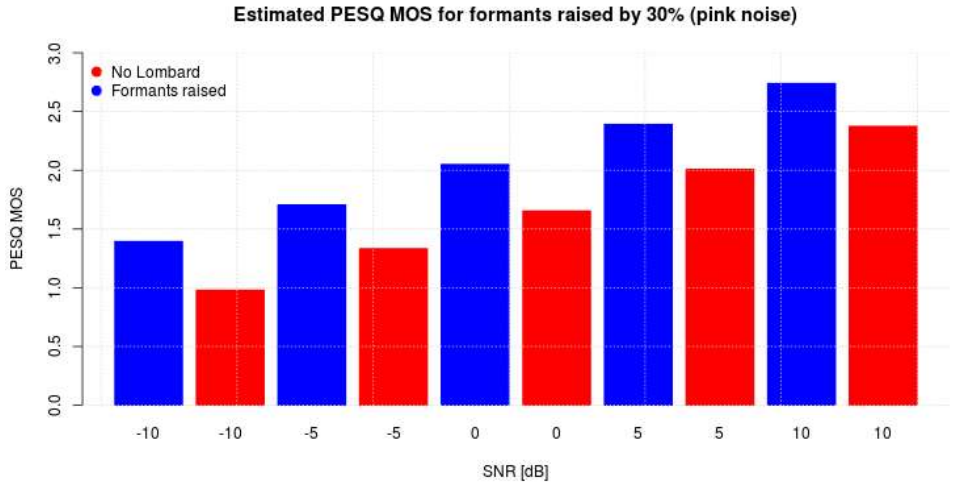



*Figure 4.2 Estimated PESQ MOS values for SNR=0, pink noise (a) and babble speech (b), denotations are as follows: CD1 – increased duration by 20%, CD2 – increased duration by 40%, formants raised: 20% (FR7) and 30% (FR8)*

a)



b)

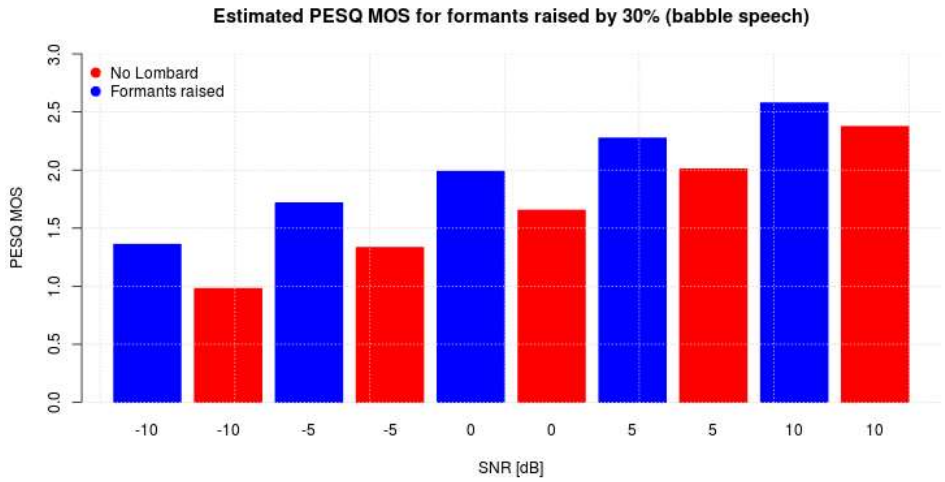



*Figure 4.3 Comparison of the PESQ MOS values of processing the most effective changes (raising formants) as a function of SNR, pink noise (a), and babble speech (b)*

In principle, modifications performed on the original speech set to increase PESQ MOS values worked with differentiated effectiveness depending on the modification type applied. However, some of the results obtained are encouraging as PESQ MOS values differ by one class between the reference and the modified signals. Contrarily, there are some cases when modification decreases PESQ MOS values.

**Classification**

In the second part of these preliminary experiments, the classification of unprocessed and processed files in noise conditions was performed. The babble speech was used as a damping factor. In this investigation, 18 Mel-Frequency Cepstral Coefficients (MFCCs) (Mermelstein, 1976) and their first-order derivatives were employed. Before feature extraction, the signal pre-processing was carried out. The speech signal was divided into 512 sample frames, and an overlap that constitutes 50% of the segment length was used. To compare classification rates, the *k*-Nearest Neighbors (kNN), a baseline machine learning algorithm, was employed. Based on obtained classification accuracies, the percent classification change was calculated:

$$change = Acc_1 - Acc_2 \tag{4.1}$$

where $Acc_1$ is the classification accuracy given for the signal after modification, and $Acc_2$ - accuracy obtained for the signal without Lombard effect.

The percent classification change is given in Table 4.2, where positive values indicate the percentage increase and negative ones – represent the percentage decrease. As seen in Table 4.2, the results are non-conclusive, they are dependent on the type of the modification used, SNR, and

the “deepness” of the interference applied. Thus, this part of the experiment will further be pursued, especially since the training set was very small.

*Table 4.2 The percentage change in classification*

| | Changes applied to the vocal tract | | | |
|---|---|---|---|---|
| | 20% to the left | 10% to the left | 10% to the right | 20% to the right |
| SNR =-10dB | 5.9 | -11.8 | -5.9 | -5.9 |
| SNR =-5dB | 0 | -17.7 | -5.9 | -5.9 |
| SNR =0dB | 5.88 | -23.5 | 0 | -23.6 |
| SNR =5dB | -35.3 | -35.3 | -41.2 | -52.9 |
| SNR =10dB | -41.2 | -11.8 | -23.6 | -41.2 |
| | Increasing the first five formants | | | |
| | 10% | by 20% | by 30% | |
| SNR =-10dB | -11.8 | 5.9 | 0 | |
| SNR =-5dB | 11.8 | 0 | 0 | |
| SNR =0dB | 17.7 | -5. | -11.8 | |
| SNR =5dB | -23.6 | -35.3 | -41.2 | |
| SNR =10dB | -11.8 | -35.3 | -23.6 | |
| | Increasing the first two formants | | | | |
| | F1 10%, F2 8% | F1 3%, F2 2% | F1 4%, F2 2% | F1 5%, F2 4% | F1 8%, F2 6% |
| SNR =-10dB | -5.9 | -11.8 | -11.8 | -11.8 | 5.9 |
| SNR =-5dB | -5.9 | 0 | 11.8 | 0 | 0 |
| SNR =0dB | 11.8 | 11.8 | 17.7 | 17.7 | 11.8 |
| SNR =5dB | -17.7 | -17.7 | -23.6 | -29.4 | -17.7 |
| SNR =10dB | -5.9 | -5.9 | -11.8 | -5.9 | -5.9 |
| | Low pass filter | | | |
| | cutoff 6000 Hz | cutoff 7000 Hz | cutoff 8000 Hz | |
| SNR =-10dB | -5.9 | -5.9 | -5.9 | |
| SNR =-5dB | -5.9 | -5.9 | -5.9 | |
| SNR =0dB | 11.8 | 0 | 0 | |
| SNR =5dB | 0 | -11.8 | -11.8 | |
| SNR =10dB | 0 | 0 | 0 | |
| | High pass filter | | | |
| | cutoff 60 Hz | cutoff 120 Hz | cutoff 180 Hz | cutoff 240 Hz |
| SNR =-10dB | 5.9 | -5.9 | -5.9 | -5.9 |
| SNR =-5dB | 0 | -5.9 | -5.9 | -5.9 |

| | | | | |
|---|---|---|---|---|
| SNR =0dB | 5.9 | -5.9 | 0 | 5.9 |
| SNR =5dB | -35.3 | -11.8 | -11.8 | -5.9 |
| SNR =10dB | 0 | 0 | 0 | 0 |
| | Change of duration | | | |
| | increase by 20% | increase by 40% | decrease by 20% | decrease by 40% |
| SNR =-10dB | -11.8 | -5.9 | 0 | 0 |
| SNR =-5dB | 0 | 0 | -5.9 | 0 |
| SNR =0dB | 5.9 | 35.3 | 5.9 | 0 |
| SNR =5dB | -17.7 | 5.9 | -23.5 | -11.8 |
| SNR =10dB | -5.9 | -5.9 | -17.7 | -5.9 |
| | Pitch smoothing | | | |
| | by 100% | by 75% | by 50% | by 25% |
| SNR =-10dB | 0 | 0 | -11.8 | -5.9 |
| SNR =-5dB | -5.9 | -5.9 | -17.7 | -11.8 |
| SNR =0dB | -17.7 | -17.7 | 11.8 | -5.9 |
| SNR =5dB | -47.1 | -53.0 | -17.7 | -29.4 |
| SNR =10dB | -35.4 | -35.3 | -29.4 | -29.4 |

## Conclusions

Performed experiments demonstrated that there is a possibility to improve speech quality by performing relatively simple operations on the input signal, judging by PESQ MOS measures. However, it was only demonstrated that the estimated Mean Opinion Score factor could be improved. The real value of MOS improvement was not verified, and it can only be examined in subjective tests. Nevertheless, the PESQ MOS algorithm is used in telecommunications in parallel with listening tests or independently due to the fact that PESQ results are correlated with subjective experiments in 93.5% (Beerends *et al.*, 2002; Rix *et al.*, 2002). One can therefore assume that the results of experiments performed with PESQ measurements may converge with the subjective examination.

A more general conclusion may be derived; namely, modifications applied are simple to obtain and may successfully be used in real-time systems working in the presence of noise, e.g., hearing aids or threat warning systems. It is also worth noticing that the additional aspect connected with Lombard speech is the problem of automatic speech recognition in noise. Based on modifications introduced, some experiments were performed and show that classification results are non-conclusive. The accuracy depends on the type of modification used, SNR, and the "deepness" of the interference applied. Thus, this part of the experiment should be further pursued, especially as a set of the training set was very small.

### 4.1.3 P.563 quality improvement – experiment 2

The main goal of the experiment was two-fold: first, to check how the speech models with the applied Lombard effect are recognizable and at what noise threshold a particular model stops working. Secondly, a quality measure is introduced, based on a feature vector derived from the signal analyzed, and then compared with the standardized metric (as described in Chapter 3.4), as well as with the MUSHRA test results. For this purpose, the models recalled in Chapter 3.1 were created utilizing all recorded speech utterances with the Lombard effect. The block diagram of the experimental setup is presented in Figure 4.4.

The following denotation of speech models are used:

M1 – harmonic model,

M2 – source-filter model with an aperiodicity parameter,

M3 – source-filter model with a waveform-based parameter,

M4 – sinusoidal model without phase preserving,

M5 – sinusoidal model with phase preserving.

Respectively, denotations for natural speech signals are the follows:

LS – utterance with the Lombard effect,

NS – original, natural speech utterance (non-Lombard).

The experiment consisted of two parts. In the first part of the experiment, an objective evaluation of models was performed. The models, as well as real speech signals, were mixed with babble speech and street noise recordings. Samples of noise were taken from the YouTube platform. The following signal-to-noise ratios (SNRs) were tested: -20 dB, -15 dB, -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB.

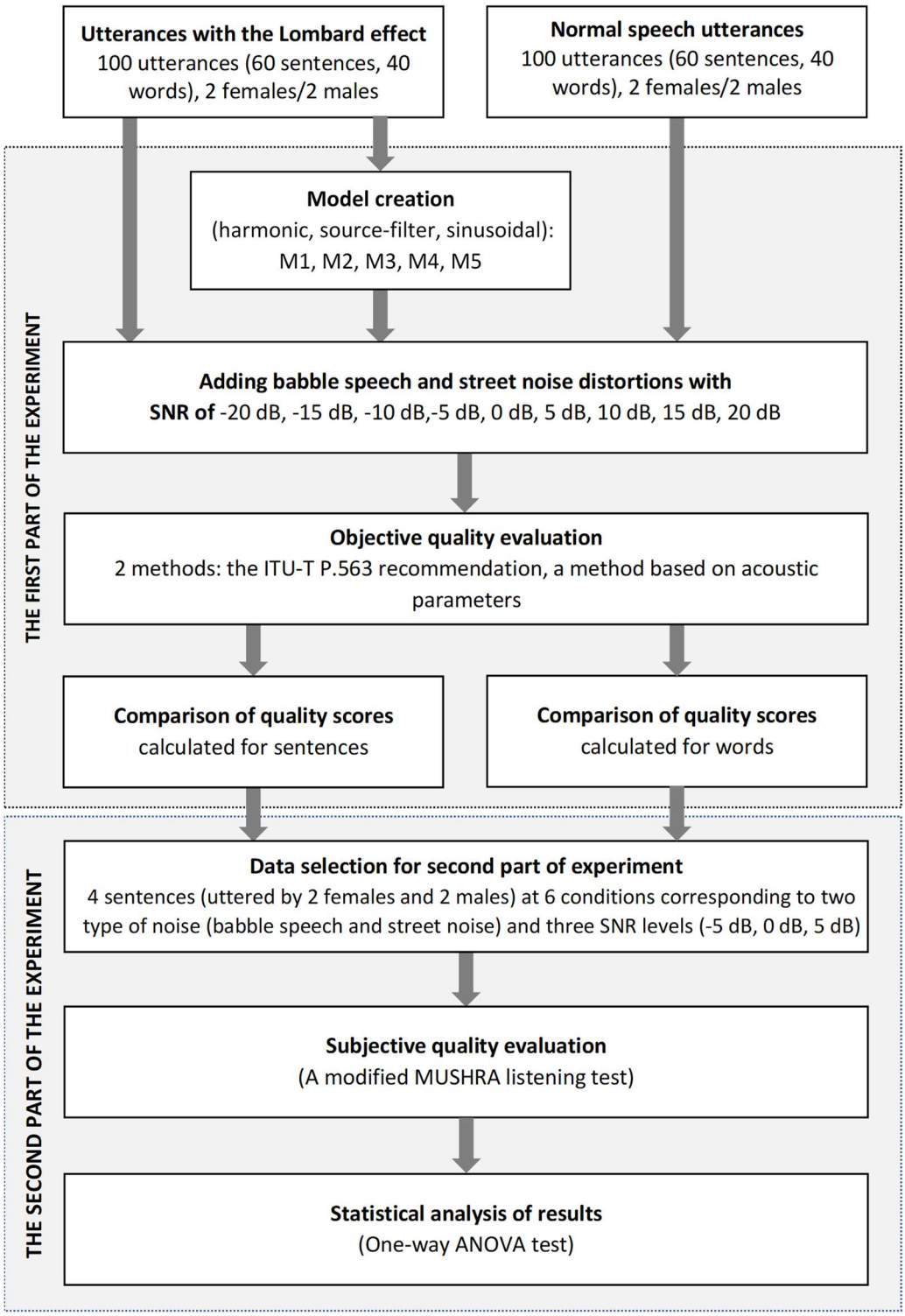



*Figure 4.4 The block diagram of the experimental setup*

**Quality evaluation techniques applied**

The quality of the models created is measured, employing both objective and subjective approaches. In this research, two objective indicators, such as P.563, defined by ITU-T recommendation (ITU-T Recommendation P.563, 2004), and a method based on acoustic parameters proposed by the author of this dissertation, are employed. The subjective quality evaluation is performed employing MUSHRA (MUltiple Stimulus with Hidden Reference and Anchors), described in the ITU-R BS. 1534-1 standard (ITU-T Recommendation BS.1534-1, 2003).

The P.563 metric was used to calculate the speech quality. It should, however, be remembered that P.563 is a single-ended measure that does not require the source (original) signal to compare. Contrarily, double-ended measures are based on the comparison of the original and the degraded signal. In the case of the experiment performed, there exists the original signal, so double-ended measures – such as PESQ (Beerends *et al.*, 2002; ITU-T Recommendation P.862, 2001) might be used. However, the applicability of double-ended measures is limited in the investigations carried out as they may not return accurate value metrics. Considering one type of noise and one type of speech modification results in four recordings to be evaluated (i.e., without noise and modification, without noise and with modification, with noise and without modification, with noise and with modification). Then, there are two possible ways of performing PESQ-based comparison:

1. The first case refers to the situation in which there is the original signal: without noise and modification, and the degraded signal contains noise, and the modification is applied – the PESQ algorithm will treat the modification of speech signal as degradation, which does not suit the aim of work.

2. The second case includes the original signal with or without modification, and the degraded signal is with or without modification, respectively, still, also with noise applied – in such conditions, speech modification is not treated as degradation. The metric will show only what the impact of noise on speech quality is. But information on how the modification potentially impacts the speech quality, which is the factor that should be measured, is not retrieved.

The above consideration shows why the double-ended measures cannot be applied to calculate the impact of modifications on the speech quality measured in the presence of noise.

**Data analyzed**

The experiments are performed on the recordings of four speakers (two males and two females). The speakers were asked to read 25 statements, including 15 sentences in Polish with different prosody (indicative, imperative, and questioning utterances) and 10 separate words.

Sentences and words used in the experiment are listed in Appendix C. These speech statements were recorded in .wav files of the audio format with the following parameters: 48 kHz; 16 bit; mono. The recording of utterances was carried out in a room with an acoustically treated interior that suppresses reverberation. The recording procedure was repeated twice: without additional noise as well as with noise interference. To simulate noise conditions, closed headphones were used. As a result, two types of recordings: 100 statements of natural, normal speech, i.e., non-Lombard speech, and 100 with the Lombard effect, were obtained.

**Result analysis**

The objective evaluation of recordings was performed separately for words and sentences. The obtained results for sentences are given in Table 4.3 and Table 4.4. Scores rated the same or higher compared to Lombard speech (LS) are highlighted in bold font.

*Table 4.3 Estimated averaged MOS-LQO values for babble speech and street noise distortions (recordings containing only sentences were used in the evaluation process)*

| | Babble speech noise | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 2.43 | 2.47 | 2.52 | 2.72 | 3.01 | 3.61 | 3.91 | 3.97 | 4.17 |
| M1 | **2.45** | 2.43 | 2.46 | 2.39 | 2.43 | 2.71 | 2.85 | 2.90 | 3.55 |
| M2 | **2.43** | 2.46 | 2.48 | **2.75** | **3.21** | **3.87** | **4.23** | **4.32** | **4.44** |
| M3 | 2.38 | 2.39 | 2.46 | **2.72** | **3.28** | **4.06** | **4.45** | **4.64** | **4.69** |
| M4 | 2.41 | 2.44 | 2.46 | 2.54 | 2.91 | 3.20 | 3.36 | 3.60 | **4.21** |
| M5 | 2.37 | 2.41 | 2.41 | 2.56 | 2.99 | 3.42 | 3.59 | 3.66 | 4.16 |
| NS | 2.41 | 2.46 | 2.42 | 2.44 | 2.67 | 3.21 | 3.49 | 3.65 | 3.70 |
| | Street noise | | | | | | | | |
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1 | 1 | 1 | 1 | 1.28 | 2.70 | 3.70 | 3.49 | 3.93 |
| M1 | 1 | 1 | 1 | 1 | 1.16 | 2.38 | 3.16 | 3.09 | 3.48 |
| M2 | 1 | 1 | 1 | 1 | 1.23 | **2.78** | **3.76** | **3.89** | **4.00** |
| M3 | 1 | 1 | 1 | 1 | **1.32** | **2.78** | 3.68 | **3.76** | **4.04** |
| M4 | 1 | 1 | 1 | 1 | **1.30** | 2.66 | 3.63 | 3.31 | 3.84 |
| M5 | 1 | 1 | 1 | 1 | **1.38** | **2.88** | **3.86** | **3.76** | **3.93** |
| NS | 1 | 1 | 1 | 1 | 1.23 | 2.32 | 3.10 | 3.23 | 3.66 |

*Table 4.4 Estimated averaged quality scores for babble speech and street noise distortions obtained by the method based on acoustic parameters derived from speech (recordings containing only sentences, not words, were used in the evaluation process)*

| | Babble speech noise | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1.13 | 1.18 | 1.52 | 2.62 | 3.58 | 4.59 | 4.92 | 4.99 | 4.98 |
| M1 | **1.29** | **1.33** | **1.54** | 2.56 | 3.50 | 4.37 | 4.62 | 4.65 | 4.62 |
| M2 | 1.12 | **1.19** | 1.52 | **2.87** | **3.82** | **4.69** | **4.95** | **5.00** | **4.99** |
| M3 | **1.15** | **1.18** | 1.46 | **2.75** | **3.72** | **4.64** | **4.94** | **4.99** | **4.98** |
| M4 | 1.12 | **1.18** | 1.33 | 2.21 | 3.02 | 4.22 | 4.59 | 4.63 | 4.59 |
| M5 | **1.16** | **1.21** | 1.49 | **2.65** | **3.61** | **4.61** | **4.92** | 4.98 | 4.96 |
| NS | 1 | 1.04 | 1.25 | 2.38 | 3.37 | 4.44 | 4.78 | 4.86 | 4.85 |
| | Street noise | | | | | | | | |
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1 | 1.06 | 1.22 | 2.18 | 3.91 | 4.74 | 4.96 | 4.98 | 4.96 |
| M1 | 1 | 1.05 | 1.17 | 2.12 | 3.64 | 4.40 | 4.58 | 4.58 | 4.56 |
| M2 | 1 | 1 | 1.19 | **2.49** | **4.13** | **4.83** | **5.00** | **5.00** | **4.98** |
| M3 | 1 | 1 | 1.18 | **2.39** | **4.03** | **4.79** | **4.98** | **4.99** | **4.97** |
| M4 | 1 | 1.05 | 1.16 | 1.88 | 3.48 | 4.38 | 4.62 | 4.60 | 4.53 |
| M5 | 1 | 1 | **1.21** | **2.20** | **3.95** | **4.77** | **4.97** | **4.98** | **4.94** |
| NS | 1 | 1 | 1.18 | **2.26** | 3.89 | 4.66 | 4.84 | 4.84 | 4.82 |

The graphical representation of results given in Table 4.3 and Table 4.4 for babble speech noise is presented in Figures 4.5 and 4.6.

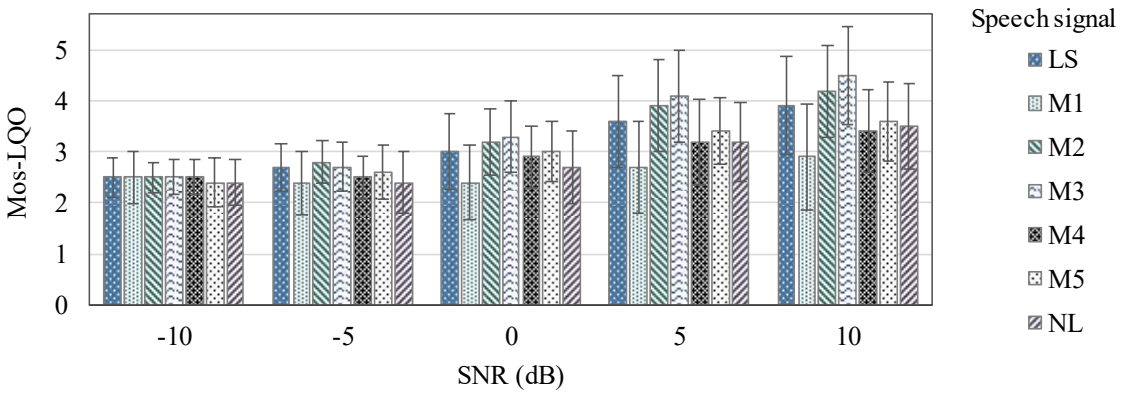



*Figure 4.5 Estimated averaged MOS-LQO values for babble speech distortions (calculated for recordings containing sentences). Denotations are as follows: speech models: M1 – harmonic model, M2 – source-filter model with aperiodicity parameter, M3 – source-filter model with a waveform-based parameter, M4 – sinusoidal model without phase preserving, M5 – sinusoidal model with phase preserving; real speech signals: LS – utterance with the Lombard effect, NS – original, natural speech utterance*

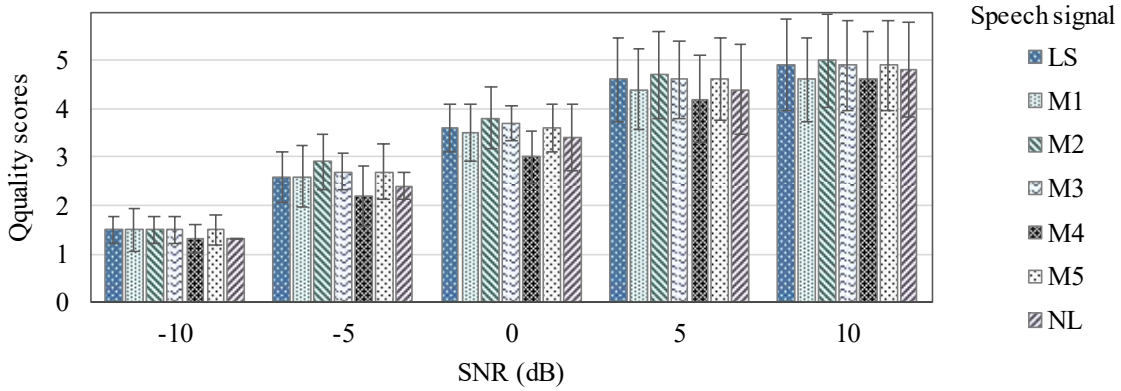



*Figure 4.6 Estimated averaged quality scores for babble speech distortions obtained by the method based on acoustic parameters (calculated for recordings containing sentences); denotations as shown in Fig. 4.5*

The graphical representation of results given in Table 4.2 and Table 4.3 for street noise is presented in Figure 4.7 and Figure 4.8.

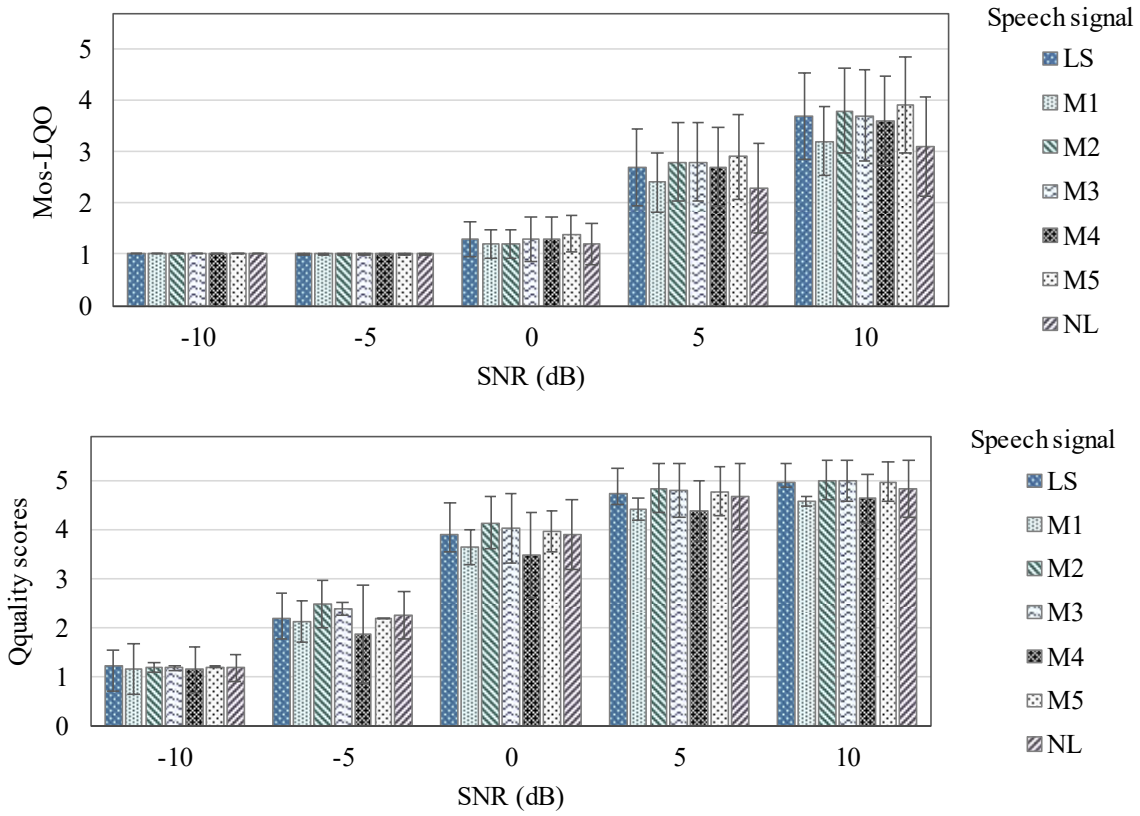



*Figure 4.7 Estimated averaged MOS-LQO values for street noise distortions (calculated for recordings containing sentences); denotations as shown in Fig. 4.5*

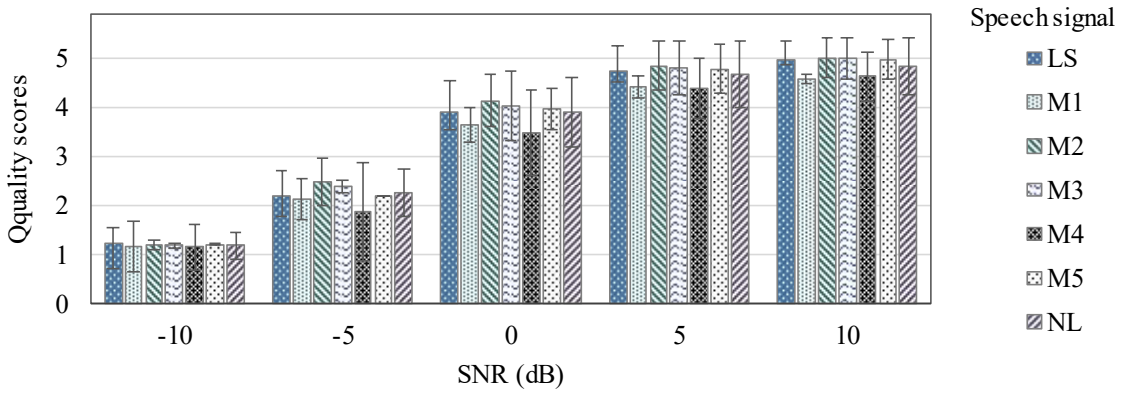



*Figure 4.8 Estimated averaged quality scores for street noise distortions obtained by the method based on acoustic parameters (calculated for recordings containing sentences); denotations as shown in Fig. 4.5*

When referring to the speech-in-noise conditions, typically, speech utterances are analyzed in the context of the Lombard effect occurrence. However, in this work, also separated words were tested to see if the Lombard effect could be applied to a single word and if it could have an impact on speech quality. The obtained results for recordings containing words are given in Table 4.5 and Table 4.6, where the scores rated the same or higher in comparison with the Lombard speech (LS) are highlighted in bold font.

*Table 4.5 Estimated averaged MOS-LQO values for babble speech and street noise distortions (recordings containing only words were used in this part of the evaluation process)*

| | Babble speech noise | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 2.27 | 2.46 | 2.65 | 3.01 | 3.63 | 4.16 | 4.14 | 4.26 | 4.28 |
| M1 | **2.29** | **2.47** | 2.58 | 2.66 | 3.15 | 3.30 | 3.35 | 3.63 | 3.71 |
| M2 | **2.30** | 2.44 | 2.64 | 2.97 | **3.77** | **4.35** | **4.46** | **4.40** | **4.44** |
| M3 | **2.41** | 2.44 | 2.58 | **3.07** | **3.83** | **4.39** | **4.54** | **4.55** | **4.54** |
| M4 | **2.32** | **2.52** | 2.54 | 2.91 | 3.55 | 3.74 | 3.82 | 3.92 | 3.92 |
| M5 | **2.29** | **2.48** | **2.65** | 2.85 | 3.51 | 3.87 | 3.70 | 3.83 | 3.73 |
| NS | **2.47** | **2.53** | **2.68** | 2.72 | 3.12 | 3.55 | 3.81 | 3.88 | 4.10 |
| | Street noise | | | | | | | | |
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1.35 | 1.36 | 1.38 | 1.24 | 1.90 | 3.40 | 4.06 | 4.02 | 4.15 |
| M1 | **1.40** | **1.41** | **1.43** | **1.33** | 1.78 | 3.08 | 3.71 | 3.75 | 3.87 |
| M2 | 1.34 | 1.35 | 1.37 | 1.23 | **1.97** | **3.67** | **4.18** | **4.05** | **4.21** |
| M3 | 1.33 | **1.36** | 1.36 | 1.23 | **1.99** | **3.71** | **4.18** | **4.16** | **4.39** |
| M4 | 1.34 | **1.36** | **1.40** | 1.21 | **2.05** | 3.34 | **4.15** | **4.11** | **4.29** |
| M5 | **1.35** | 1.35 | 1.37 | 1.24 | **2.00** | **3.43** | 3.82 | 3.70 | 3.80 |
| NS | **1.53** | **1.53** | **1.54** | **1.48** | 1.81 | 3.00 | 3.52 | 3.54 | 3.92 |

*Table 4.6 Estimated averaged quality scores for babble speech and street noise distortions obtained by the method based on acoustic parameters (recordings containing only words were used in this part of the evaluation process)*

| | Babble speech noise | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1.29 | 1.32 | 1.60 | 2.44 | 3.12 | 4.34 | 4.82 | 4.96 | 4.99 |
| M1 | 1.23 | 1.24 | 1.47 | 2.18 | 2.70 | 3.82 | 4.14 | 4.21 | 4.20 |
| M2 | **1.32** | **1.37** | **1.70** | **2.49** | **3.49** | **4.45** | **4.86** | **4.97** | **5.00** |
| M3 | 1.28 | 1.30 | **1.68** | **2.52** | **3.33** | **4.41** | **4.83** | 4.92 | 4.93 |
| M4 | 1.28 | 1.28 | 1.41 | 2.12 | 2.57 | 3.77 | 4.19 | 4.23 | 4.19 |
| M5 | 1.26 | 1.23 | 1.40 | 1.95 | 2.12 | 3.30 | 3.73 | 3.80 | 3.78 |
| NS | 1.00 | 1.07 | 1.35 | 2.37 | **3.17** | 4.12 | 4.43 | 4.50 | 4.50 |
| | Street noise | | | | | | | | |
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1 | 1 | 1 | 2 | 3.62 | 4.55 | 4.89 | 4.97 | 4.99 |
| M1 | 1 | 1 | 1 | 2 | 3.40 | 4.24 | 4.53 | 4.57 | 4.57 |
| M2 | 1 | 1 | 1 | 2 | **3.76** | **4.62** | **4.93** | **5.00** | **5.00** |
| M3 | 1 | 1 | 1 | 2 | **3.74** | **4.61** | **4.90** | 4.96 | 4.96 |
| M4 | 1 | 1 | 1 | 2 | 3.21 | 4.18 | 4.52 | 4.57 | 4.55 |
| M5 | 1 | 1 | 1 | 2 | 2.98 | 3.91 | 4.30 | 4.37 | 4.36 |
| NS | 1 | 1 | 2 | 2 | **3.77** | 4.49 | 4.71 | 4.74 | 4.73 |

According to the grading scale, the quality of sounds that achieves the approximated score equal to 3 refers to “fair” quality, and they may be considered slightly annoying. Based on this result, the answer may be given as to what SNR level threshold a particular model stops working. It occurs that this threshold is -5 dB in the case of babble speech noise for both objective quality evaluation techniques. For street noise distortions, thresholds are 5 dB and 0 dB with respect to the P.563 indicator and method based on acoustic parameters, respectively.

Based on the objective results of the word model evaluation (see Table 4.5 and Table 4.6), the same SNR level thresholds were established as in the case of the sentence model assessment (i.e., -5 dB in the case of babble speech noise in addition to 5 dB and 0 dB for street noise distortions).

It is worth noting that with the addition of babble noise of very high volume, MOS values at SNRs at -20 dB, -15 dB, or -10 dB indicate that the sound quality is good enough (see Tables 2, 4). However, these results are not reliable because the estimated LS values are lower than NS values. In fact, the opposite is true, i.e., the LS values at high noise levels should be higher than those of NS. This may be caused by the fact that the added babble noise contains speech, and the quality ratings obtained refer to noise rather than the signal.

The objective measures show that in most cases, the best scores are achieved with both source-filter models and the model based on the sinusoids with phase preserving. Contrarily, the measure based on parameterization shows a smaller difference between the models than in the case of the P.563 indicator values. A listening test should be performed to check whether the proposed measure overestimates the quality of models or MOS-LQO underestimates it.

When comparing the obtained results to the state-of-the-art, one can see that such a comparison in practice is not straightforward. For example, Michelsanti et al. (Michelsanti *et al.*, 2019) reported averaged scores of PESQ and ESTOI (short-time objective intelligibility) (Jensen and Taal, 2016) measures for the deep-learning-based system of audio-visual speech enhancement with the Lombard effect applied. To elicit the Lombard effect, speech-shaped noise (SSN) at 80 dB sound pressure level (SPL) was presented to the speakers while they were reading the sentences (Michelsanti *et al.*, 2019). It is worth noting that ESTOI scores, which estimate speech intelligibility, range from 0 to 1, where high values correspond to high speech intelligibility. When trained on a narrow SNR range, for the audio-only case with the Lombard effect (AO-L), PESQ measurement returned a value of 1.283, and ESTOI was equal to 0.448. Contrarily, when the system was trained on a wide SNR range, the averaged values ranged between 1.346 (for -20 to 5 dB) to 3.127 (for 10 to -30 dB) for the AO-L case. ESTOI values changed dramatically from 0.442 for -20 dB to -5 dB SNR up to 0.927 for the SNR range between 10 and 30 dB. So, the relative performance of the systems at SNR $\leq$ 5 dB is similar to the one observed for the systems trained on a narrow SNR range (Michelsanti *et al.*, 2019).

As seen from this discussion, not only the experimental setup was different compared to the approach presented, but also the analysis differed from the one performed within this dissertation work; thus, a direct comparison is not possible. Even the observation on at what SNR value the model does not work seems not to be common; in the case of this research, the models stop working at the threshold of -5 dB, in the work of Michelsanti et al. (Michelsanti *et al.*, 2019), it refers to 5 dB.

**Subjective test results**

An informal listening test showed that the quality of the Lombard speech models might be directly compared to the original sound. Therefore, in the second part of the experiment, the

subjective quality evaluation is carried out. In the listening experiment, the participants compared the performance of different models to the natural utterances of Lombard speech in noise. To ensure that the test session does not take more than 20 minutes, four Lombard speech utterances consisting of sentences uttered by four speakers were used. Also, only the three models, which showed the best results in the first part of the experiment (M2, M3, and M5) and the Lombard speech utterances, were evaluated. Because of the time constraint, only two types of noise recordings (babble speech and street noise) were mixed with speech models. Besides, the following signal-to-noise ratio (SNR) was considered: -5 dB, 0 dB, and 5 dB. As a result, six test conditions corresponding to the combination of noise type and SNR were used in the listening test. The average duration of the MUSHRA test session was approx. 19 minutes. Twelve speech processing experts from the Vilnius University Institute of Data Science and Digital Technologies and the Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics took part in this test. The obtained results are given in Table 4.7.

The subjective test results contained in Table 4.7 show that original recordings of the Lombard speech are more intelligible in the noise condition than their models (except for one example, which is highlighted in bold font). A visualization of the subjective test results is given in Figure 4.9 and Figure 4.10.

*Table 4.7 Subjective quality scores for babble speech and street noise distortions; denotations are as follows: real speech signals: LS – utterance with the Lombard effect, speech models: M2 – source-filter model with aperiodicity parameter, M3 – source-filter model with a waveform-based parameter, M5 – sinusoidal model with phase preserving*

| | Babble speech noise | | | Street noise | | |
|---|---|---|---|---|---|---|
| SNR | –5 dB | 0 dB | 5 dB | –5 dB | 0 dB | 5 dB |
| LS | 64.11 | 66.91 | 76.34 | 28.39 | 48.55 | 64.89 |
| M2 | 64.05 | 65.27 | 70.39 | 31.43 | 45.07 | 61.34 |
| M3 | 62.75 | 63.39 | 66.32 | **28.41** | 43.73 | 60.68 |
| M5 | 52.61 | 50.77 | 53.77 | 31.59 | 46.48 | 60.20 |

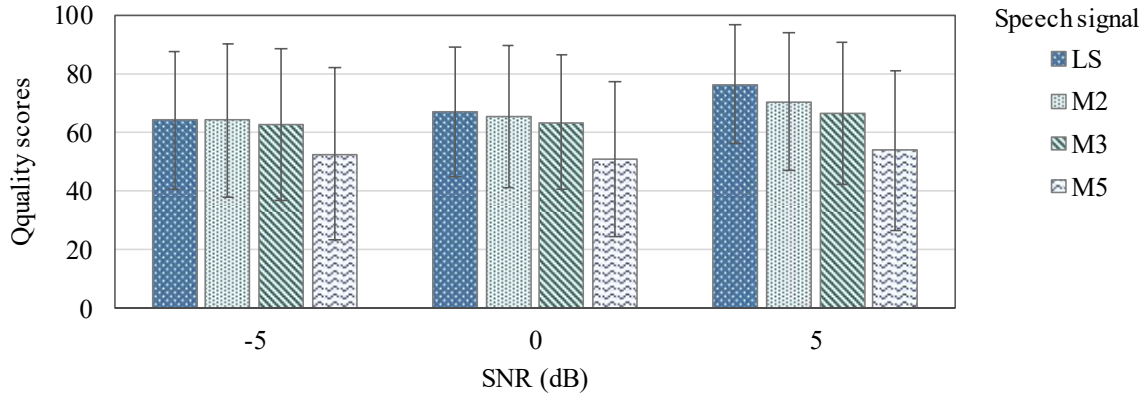



*Figure 4.9 Subjective quality scores for babble speech noise; denotations as shown in Table 4.7*

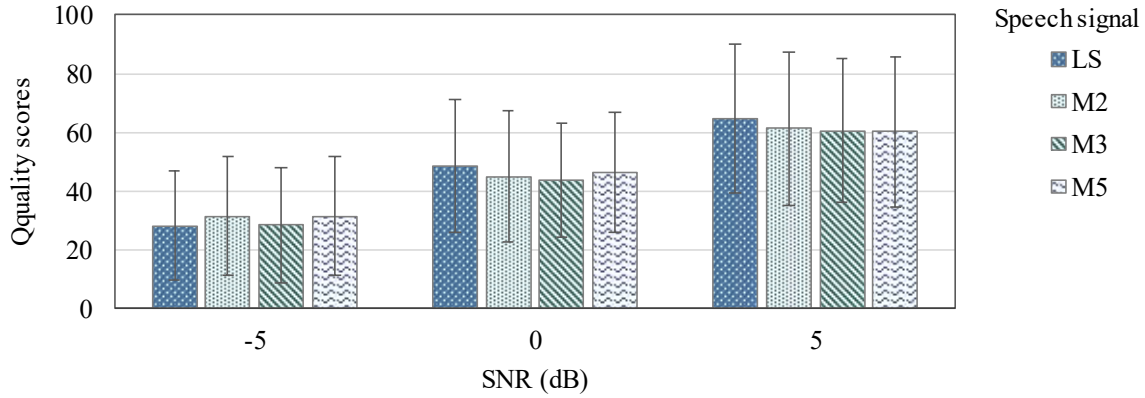



*Figure 4.10 Subjective quality scores for street noise; denotations as shown in Table 4.7*

In line with the earlier shown results (see Figure 4.9 and Figure 4.10), one can observe that the SNR level threshold at which a particular model stops working is 0 dB in the case of street noise. Contrarily, for babble speech noise, it is not possible to determine such a threshold based on the experiment carried out.

Results obtained by Michelsanti et al. (Michelsanti *et al.*, 2019) refer to two types of subjective tests, namely MUSHRA and the speech intelligibility test. For the AO-L at -5 dB SNR, the result was approx. 25 points, whereas for 5 dB SNR returned approx. 50 points. Obviously, the intelligibility test also depends on the SNR. Moreover, it was tested for several types of words, i.e., color, letter, and digit. The mean intelligibility scores are within the range of approx. 35% for -20 dB SNR to approx. 85% for 5 dB SNR. Again, all analysis conditions differ from the ones performed by us, thus, a straightforward comparison is not possible. However, MUSHRA scores for street noise are low for – 5dB SNR, and they are at the same level as in work by Michelsanti

et al. (Michelsanti *et al.*, 2019). Contrarily, they are higher for SNR equals 5dB for both street noise and babble speech conditions.

Seshadri et al. (Seshadri *et al.*, 2019) reported MUSHRA-based scores when applying the Lombard effect to several vocoders. To induce Lombard speech, background noise nonstationary pub noise [60], with an A-weighted sound pressure level (SPL) of approximately 80 dB), was presented to the speakers' ears with headphones while they were being recorded (Seshadri *et al.*, 2019). Scores were shown for parametric vocoders (VOCs) for feature extraction and machine learning models (MLMs) for speech modifications. The MUSHRA test aimed at evaluating the "Lombardness" of the utterances from different VOC and MLM combinations of a single sentence (same speaker and linguistic content). All results were conditioned by various vocoders employed. The scores ranged between approx. 40 to 60 points of the mean "Lombardness." Moreover, the CMOS quality test was applied as well as the so-called instrumental intelligibility test, given in bits/s. Also, in this case, it is not possible to directly compare the results reported by Seshadri et al. and the results obtained in the experiments carried out.

**Statistical results**

To check whether the differences between the measurements are statistically significant, a statistical analysis of the results is performed. For this purpose, the one-way ANOVA test was employed, which is used to measure the variation between the utterance with the Lombard effect (LS) and their models. The null hypothesis ($H_0$) states that the utterance and its model are from populations with the same means. The decision rule to reject this hypothesis can be expressed by the following formula:

$$reject\ H_0\ if\ F > F_{critical}(1-\alpha) \tag{4.2}$$

where $F$ is the calculated test statistic and $F_{critical}$ is the critical value taken from the $F$-distribution table. Details on how to perform analyses using ANOVA as well as critical values of $F$-distribution, are given in the textbook of Tabachnick and Fidell (Tabachnick and Fidell, 2007). The test significance level equals 0.05 (based on the ITU-R Recommendation BS.1534-1 (ITU-T Recommendation BS.1534-1, 2003)).

The test results are given in Tables 4.8-4.10. Differences that are statistically significant (see Table 4.8 and Table 4.10) are highlighted in bold font.

*Table 4.8 The result of the ANOVA test (F-values) for MOS-LQO quality scores*

| | Babble speech noise | | | Street noise | | |
|---|---|---|---|---|---|---|
| SNR | –5 dB | 0 dB | 5 dB | –5 dB | 0 dB | 5 dB |
| M2 | 0.1306 | 2.2696 | 2.8214 | – | 0.62198 | 0.31311 |
| M3 | 0.0004 | **4.1235** | **8.1662** | – | 1.05374 | 0.28167 |
| M5 | 2.6309 | 0.0187 | 1.3709 | – | 1.96677 | 1.67492 |

*Table 4.9 The result of the ANOVA test (F-values) for quality scores obtained by the method proposed*

| | Babble speech noise | | | Street noise | | |
|---|---|---|---|---|---|---|
| SNR | –5 dB | 0 dB | 5 dB | –5 dB | 0 dB | 5 dB |
| M2 | 0.0001 | 0.0675 | 0.0000 | 0.0431 | 0.0231 | 0.0039 |
| M3 | 0.0047 | 0.0047 | 0.0008 | 0.0199 | 0.0070 | 0.0012 |
| M5 | 0.0013 | 0.0015 | 0.0002 | 0.0001 | 0.0008 | 0.0003 |

*Table 4.10 The result of the ANOVA test (F-values) for subjective quality scores*

| | Babble speech noise | | | Street noise | | |
|---|---|---|---|---|---|---|
| SNR | –5 dB | 0 dB | 5 dB | –5 dB | 0 dB | 5 dB |
| M2 | 0.0002 | 0.1077 | 1.4978 | 0.67742 | 0.53348 | 0.40766 |
| M3 | 0.0607 | 0.5462 | **4.1795** | 0.00311 | 1.21187 | 0.6159 |
| M5 | 3.6195 | **8.8708** | **17.8488** | 0.70643 | 0.23492 | 0.69137 |

## 4.2 Lombard speech detection process

Speech type detection is one of the critical elements of the adaptive system, proposed and discussed in this dissertation. Speech can be characterized by many parameters presented in the previous chapters, but the process of detection must be simple and fast enough to be employed in the near real-time process.

Speech type detection may be treated as a binary classification problem – in other words, speech can be "Lombard" or "non-Lombard." Of course, in reality, the speech signal is way more complicated than just this simple Lombard vs. non-Lombard differentiation. There are moments of silence, non-sounding fragments, etc. Therefore, the classification is always an approximation of the speech type, but it allows for better speech modification algorithm selection.

Figure 4.11 presents the topics that this Chapter will cover in the context of the experiments performed.

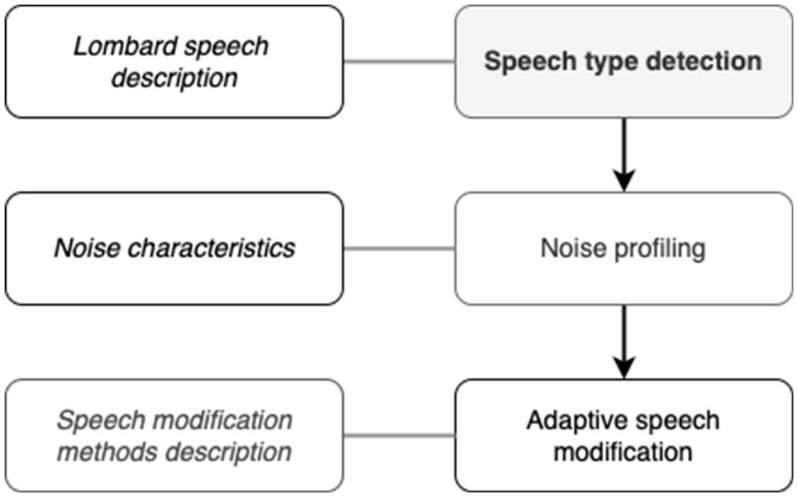



*Figure 4.11 Structure of the dissertation with current Chapter topics highlighted*

The idea of this process is to recognize the speech type using a typical classification model with two output classes – Lombard or non-Lombard. Various 2D representation techniques have been verified to identify the best approach.

The general idea is as follows: since the speech signals are analyzed in near real-time conditions, the visualizations might help detect the character of the given moment of speech. Assuming that there is a neural network that detects the type of the speech with high accuracy based on the speech visualization, one can use this classification to apply it in a decision pipeline

later (i.e., whether to modify the speech further or not – because it already has the Lombard effect-like features).

The assumptions adopted are shown in Figure 4.12.

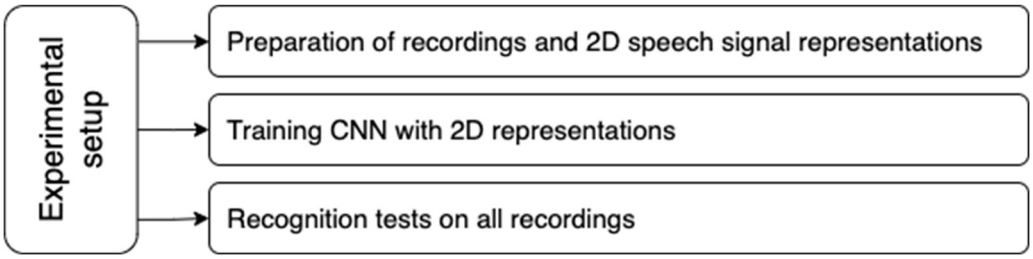



*Figure 4.12 Experimental setup*

### 4.2.1 Assumptions

As it was already discussed, one of the most important elements of speech signal processing is speech type detection – whether it is neutral or Lombard speech. Of course, it's quite often not very specific and exact differentiation, but it is crucial from the point of view of the further processing.

For instance, if one of the elements in the process would be increasing fundamental frequency (F0) and the signal which will be processed is already Lombard-type (so F0 rises naturally), the processing will lead to speech quality degradation and loss of its intelligibility. That is why detecting the speech type is so crucial in the system.

For the Lombard speech recognition algorithm to be used in real-time systems (e.g., in broadcasting systems), it is necessary to ensure the operation of such a system in real-time. Processing algorithms for changing neutral speech into Lombard cannot be too computationally complex, nor should it cause a longer delay in the analysis and the processing itself. For the experiments in this work, it was assumed that a delay of 0.5-0.7 seconds is acceptable. Consecutively, inference (recognition of the type of speech) must be performed almost in real-time, preferably without delays. Therefore, the models of neural networks presented are relatively simple − the more complex the model, the more time it takes to perform the analysis, which in turn would result in delays. The hardware on which the processing and recognition are performed has a major impact on inference speed. Therefore, for this work, the possibility of making inferences in real-time was not investigated; it was only assumed that the process of recognizing a single recording should take no more than the recording itself (performance on a personal computer).

At the same time, due to the low reliability of systems based on the analysis of the physical characteristics of the speech signal and the need to have a comparative signal during detection, the focus was on methods based on deep learning. It was also assumed that the very process of training the neural network might be long - it does not affect the speed of recognizing the type of speech signal in any way.

An additional assumption was the availability of sufficiently short but varied speech signals for training. It was assumed that only short sentences with variable prosody are suitable for training. No separate words were used for training because their recording takes place in artificial conditions, i.e., the person being recorded reads the list of words. The point, however, is that under normal conditions, words are part of a larger whole, and the lack of use of natural prosody could significantly limit the differentiation of speech signal characteristics, very important from the point of view of training the neural network.

### 4.2.2 Preparation of recordings

For the purposes of training and inference, two sets of recordings in two languages were used:

- Recordings in German - several seconds-long statements (40 sentences), eight speakers, including three women and five men. Each sentence was recorded under conditions of silence and with accompanying disturbance (Soloducha *et al.*, 2016).
- Recordings in Polish - several seconds-long statements (15 sentences), four speakers, including two women and two men. Each sentence was recorded under conditions of silence and with accompanying disturbance (Czyzewski *et al.*, 2017).

All sets contain recordings of neutral and Lombard speech, thanks to which it is possible to segment the recordings, label them and use them in the network training in the supervised learning process.

The process of preparing recordings includes the following steps:

- Calculating STFT and amplitude value, the length of the window is 512 samples. Hop length is set to half of the window length.
- The next step is to truncate the first 10 spectrum values. They do not really carry any information, and they add a lot of noise to the spectrogram.
- The next step is to remove all voiceless fragments from the spectrum, i.e, those where there is essentially no energy. The effect is that if less than 90% of the content of a given window does not carry information about the speech signal (it is voiceless or simply silent or disturbed), then such a window is not included in the training. The point is that the silence window can misclassify a given type of speech.

- The last step is to generate the visualization and save it as an image (png), scaled to such a resolution that it is effective for the training algorithm (too high resolution causes the necessity to use a large amount of memory and extends the learning process, while not bringing improvement).

The following 2D representations were used in this work:

- Spectrogram;
- Chromagram;
- Mel spectrogram;
- MFCC without rescaling;
- MFCC rescaled.

All 2D representations are resized to a resolution of 90x93 pixels in 4 channels (RGBa). Each image refers to about 0.25 seconds of recording, and each recording is the source of many pictures (how many – it depends on the content of the information about the speech signal). For example, for German recordings, 40 sentences, eight speakers, and two types (in silence and noise) are a total of 99pprox.. 5,000 saved images for training. The number of these pictures depends on the criteria for deleting images without speech, pitch, and window length.

All saved images are labeled with the gender of the speaker and whether or not there was noise during the recording of the speaker.

### 4.2.3 Experiments

Several models of convolutional neural networks (CNN – Convolutional Neural Network) were used in the training process. Many experiments were performed to show which network model was the most effective. Simple network models were used to optimize learning time in relation to outcomes. These experiments are numbered from 1 to 9 (see Fig. 4.13).

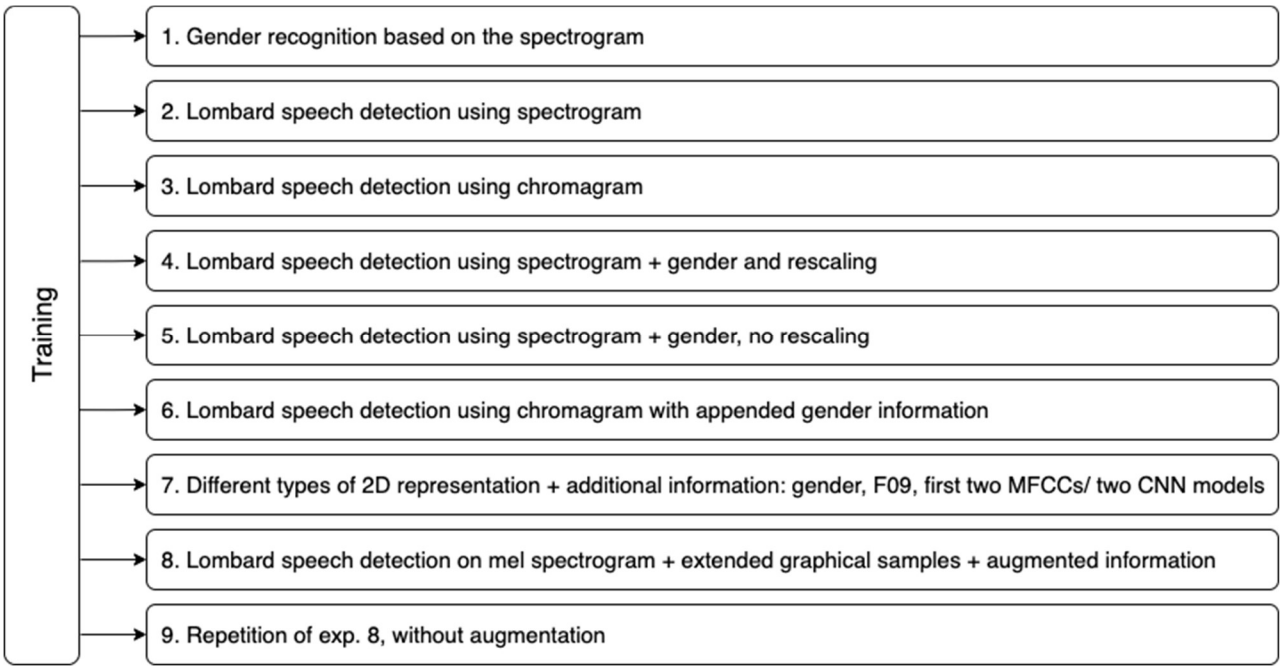



*Figure 4.13 Training experiments*

Models of the convolutional neural networks are presented in a tabular format, describing all layers and transformations. The layers description is done using the following names and shortcuts:

- Conv2d – is a basic two-dimensional convolutional layer (a two-dimensional convolutional layer means that the input matrix is 3-dimensional – representing width, height, and the number of filters).
- Max_pooling2d – is a max-pooling layer, reducing the dimensions of the input layer.
- Dropout – is an operation of randomly ignoring selected neurons in the learning process; this is a method of regularization and thus improves the results of learning in terms of the ability to generalize.
- Flatten – this is a flattening operation which means that the 3-dimensional output matrix is flattened into a vector that can be used in the typical, dense layer.
- Dense – is a basic neural network flat layer.

The set of 2D representations used for training and validation is divided as follows:

- 2/3 of the whole set is used for training, out of which 7% of 2D representations are used for accuracy verification.
- 1/3 of the whole set is a testing set, not taking part in the learning process at all.

**Gender recognition based on the spectrogram (experiment 1)**

This experiment is only an introductory step in recognition of Lombard speech, as it has been hypothesized that information about gender may be a vital feature supporting the recognition process.

The model of the Convolutional Neural Network used is presented in Table 4.11. This description shows the number of filters used in every convolutional layer: the first layer contains 32 filters, the second one 16 filters.

*Table 4.11 Model of the network used in experiment 1*

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_2 (Conv2D) | (None, 90, 93, 32) | 544 |
| max_pooling2d_2 (MaxPooling2 | (None, 45, 46, 32) | 0 |
| dropout_3 (Dropout) | (None, 45, 46, 32) | 0 |
| conv2d_3 (Conv2D) | (None, 45, 46, 16) | 2064 |
| max_pooling2d_3 (MaxPooling2 | (None, 22, 23, 16) | 0 |
| dropout_4 (Dropout) | (None, 22, 23, 16) | 0 |
| flatten_1 (Flatten) | (None, 8096) | 0 |
| dense_2 (Dense) | (None, 256) | 2072832 |
| dropout_5 (Dropout) | (None, 256) | 0 |
| dense_3 (Dense) | (None, 2) | 514 |

**Accuracy in the testing set: 93%**

Sample gender recognition results on the testing set are presented in Figure 4.14.

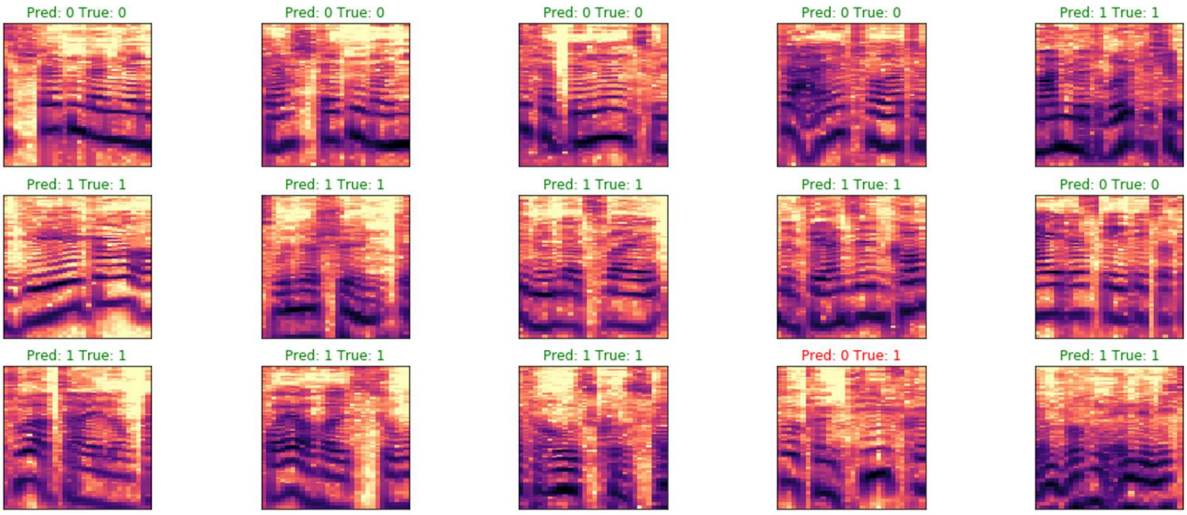



*Figure 4.14 Gender recognition results. Pred – predicted value, True – true value, 1 – male, 0 – female. Red descriptions mean wrong recognition results*

**Lombard speech detection using spectrogram (experiment 2)**

The same set of 2D representations has been used to train the network with the goal of recognizing the Lombard speech type. In other words, it was a two-classes classification problem. The model used to train this recognition challenge is presented in Table 4.12.

*Table 4.12 Model of CNN used in experiment 2*

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_2 (Conv2D) | (None, 90, 93, 32) | 544 |
| max_pooling2d_2 (MaxPooling2 | (None, 45, 46, 32) | 0 |
| dropout_3 (Dropout) | (None, 45, 46, 32) | 0 |
| conv2d_3 (Conv2D) | (None, 45, 46, 16) | 2064 |
| max_pooling2d_3 (MaxPooling2 | (None, 22, 23, 16) | 0 |
| dropout_4 (Dropout) | (None, 22, 23, 16) | 0 |
| flatten_1 (Flatten) | (None, 8096) | 0 |

| | | |
|---|---|---|
| dense_2 (Dense) | (None, 256) | 2072832 |
| dropout_5 (Dropout) | (None, 256) | 0 |
| dense_3 (Dense) | (None, 2) | 514 |

**Accuracy in the testing set:** 76%

Sample classification results are presented in Figure 4.15.

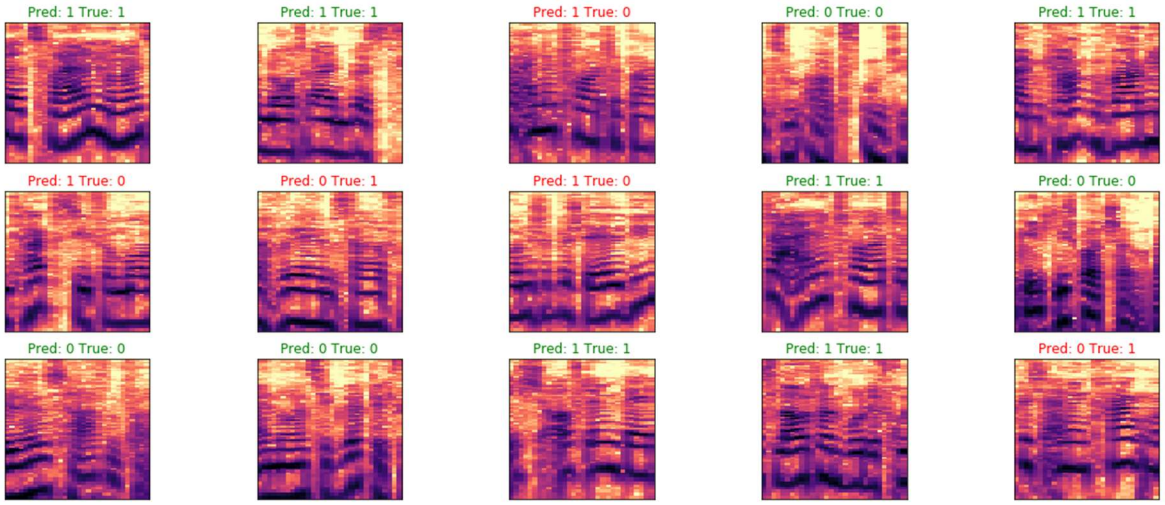



*Figure 4.15 Sample classification results using the CNN trained in experiment 2. Pred – predicted result, True – true label, 1 – Lombard speech, 0 – neutral speech*

The accuracy is not satisfying, and the sample dataset shows multiple wrong recognition results.

**Lombard speech detection using chromagram (experiment 3)**

In this experiment, the concept is similar to the previous one, with a different visualization selected – chromagram. The implemented model is presented in Table 4.13.

*Table 4.13 CNN model for experiment 3*

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_2 (Conv2D) | (None, 90, 93, 32) | 544 |

| | | |
|---|---|---|
| max_pooling2d_2 (MaxPooling2 | (None, 45, 46, 32) | 0 |
| dropout_3 (Dropout) | (None, 45, 46, 32) | 0 |
| conv2d_3 (Conv2D) | (None, 45, 46, 16) | 2064 |
| max_pooling2d_3 (MaxPooling2 | (None, 22, 23, 16) | 0 |
| dropout_4 (Dropout) | (None, 22, 23, 16) | 0 |
| flatten_1 (Flatten) | (None, 8096) | 0 |
| dense_2 (Dense) | (None, 256) | 2072832 |
| dropout_5 (Dropout) | (None, 256) | 0 |
| dense_3 (Dense) | (None, 2) | 514 |

**Accuracy in the testing set:** 58%

Sample classifications are presented in Figure 4.16. The accuracy is too low to treat it as a promising alternative.

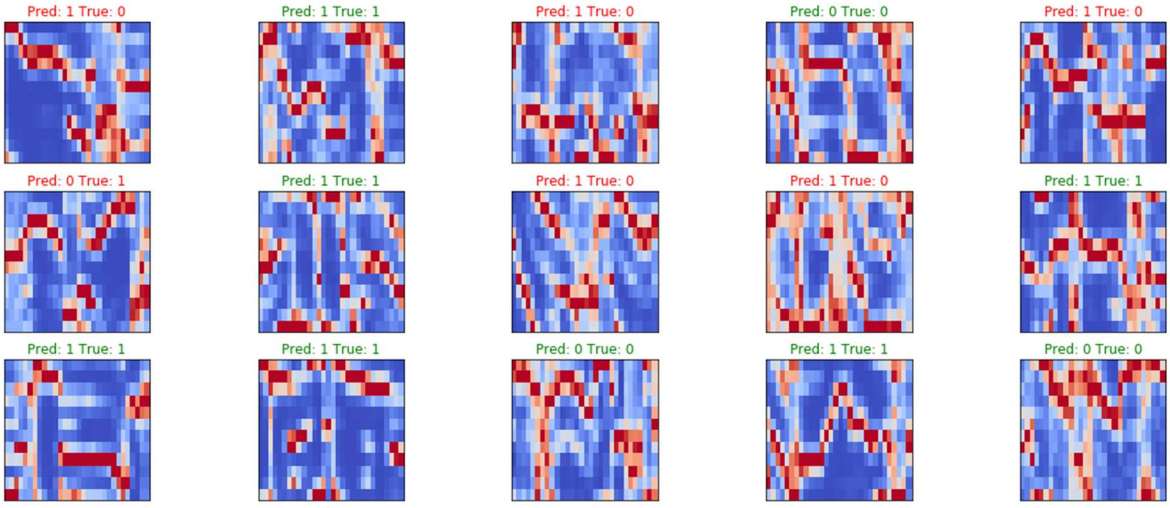



*Figure 4.16 Sample classification results using the CNN trained in experiment 3. Pred – predicted result, True – true label, 1 – Lombard speech, 0 – neutral speech*

**Lombard speech detection using spectrogram with appended information about the gender and with rescaling (experiment 4)**

In this experiment, the alpha channel has been replaced with the gender bit. Of course, every pixel in a „normal" picture is stored as a 4-byte information component (3 bytes for colors and 1 byte for alpha channel), and later in the process of training, every byte is rescaled to the range 0 to 1. This means that – because gender might be 0 or 1, the rescaled values of gender might be 0 or 1/255. It will probably have then little effect on the learning process.

The CNN model is slightly changed in this experiment – the last dense layer has 512 neurons. The model is presented in Table 4.14.

*Table 4.14 CNN model used in experiment 4*

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_4 (Conv2D) | (None, 90, 93, 32) | 544 |
| max_pooling2d_4 (MaxPooling2 | (None, 45, 46, 32) | 0 |
| dropout_6 (Dropout) | (None, 45, 46, 32) | 0 |
| conv2d_5 (Conv2D) | (None, 45, 46, 16) | 2064 |
| max_pooling2d_5 (MaxPooling2 | (None, 22, 23, 16) | 0 |
| dropout_7 (Dropout) | (None, 22, 23, 16) | 0 |
| flatten_2 (Flatten) | (None, 8096) | 0 |
| dense_4 (Dense) | (None, 512) | 4145664 |
| dropout_8 (Dropout) | (None, 512) | 0 |
| dense_5 (Dense) | (None, 2) | 1026 |

**Accuracy on the testing set**: 80%

Sample classifications are presented in Figure 4.17.

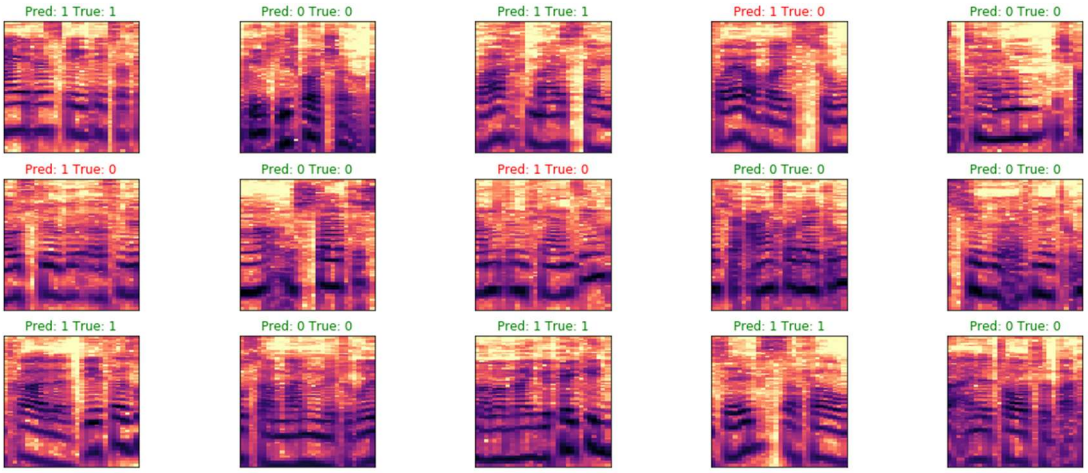



*Figure 4.17 Sample classification results using the CNN trained in experiment 4. Pred – predicted result, True – true label, 1 – Lombard speech, 0 – neutral speech*

**Lombard speech detection using spectrogram with appended gender information without rescaling (experiment 5)**

In this experiment, the alpha channel was replaced with the gender bit. Of course, every pixel in a "normal" picture is stored as a 4-byte component (3 bytes for color information, 1 byte for alpha channel). Later in the experiment, every byte value is rescaled to the range 0-1. In this experiment, the impact of the gender bit was increased by setting its value to either 64 or 192, which means that after rescaling, its value is 0.25 or 0.75.

The CNN model used is identical to experiment 4. Sample classifications are presented in Figure 4.18.

**Accuracy on the training set:** 82.5%

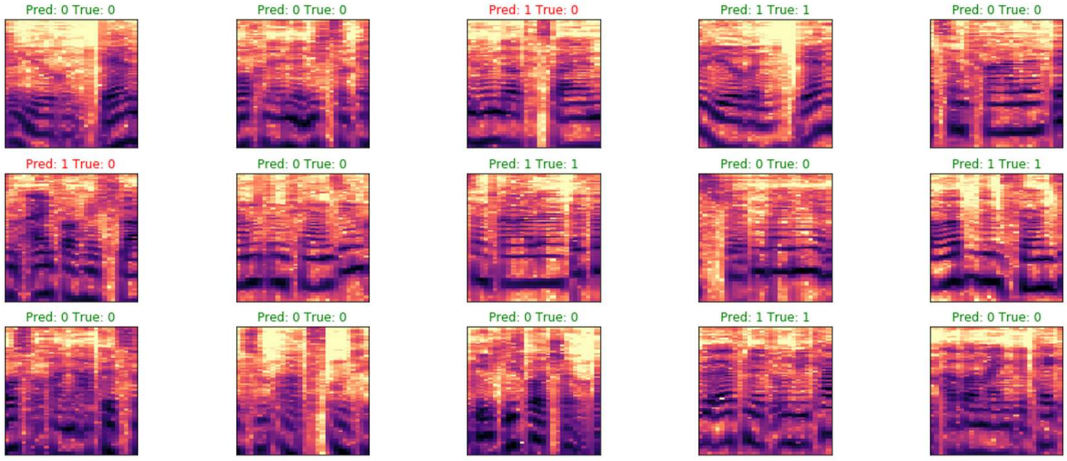



*Figure 4.18 Sample classification results using the CNN trained in experiment 5. Pred – predicted result, True – true label, 1 – Lombard speech, 0 – neutral speech*

**Lombard speech detection using chromagram with appended gender information (experiment 6)**

Gender information was appended similarly as in experiment 5.

**Accuracy on the testing set: 66%**

Sample classifications are presented in Figure 4.19.

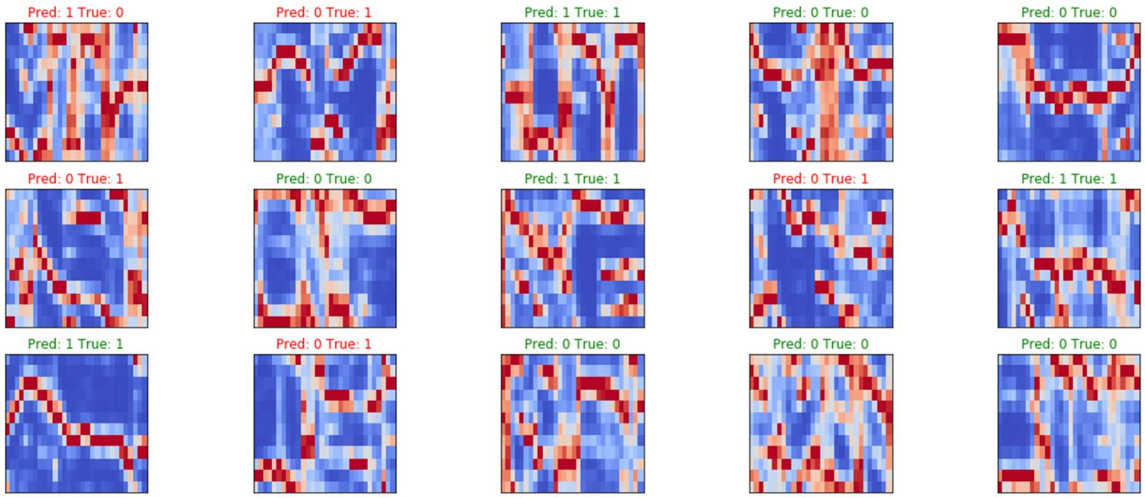


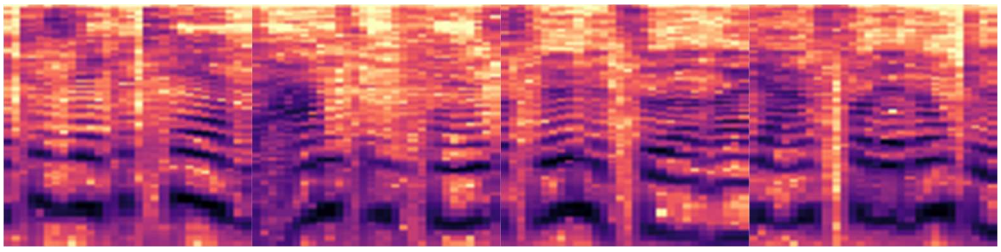

*Figure 4.19 Sample classification results using the CNN trained in experiment 6. Pred – predicted result, True – true label, 1 – Lombard speech, 0 – neutral speech*

**Comparison of the different types of visualization and recognition performance (experiment 7)**

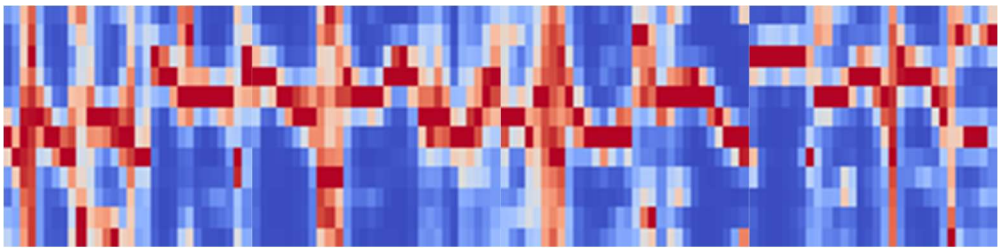
Since previous experiments showed that it is crucial to select the appropriate graphical representations and to augment the data properly, the following approach involved testing different graphical representations and their effectiveness with comparable models and the same training time.

Various graphical representations, including a short fragment of the speech recording (approx. 0.5 seconds), have been tested. The graphics have been presented in Figure 4.20.

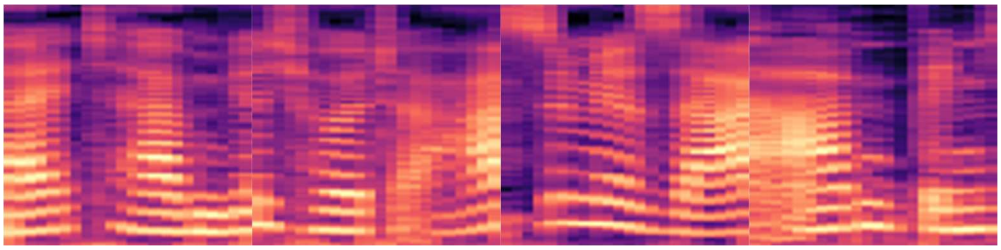
Spectrogram



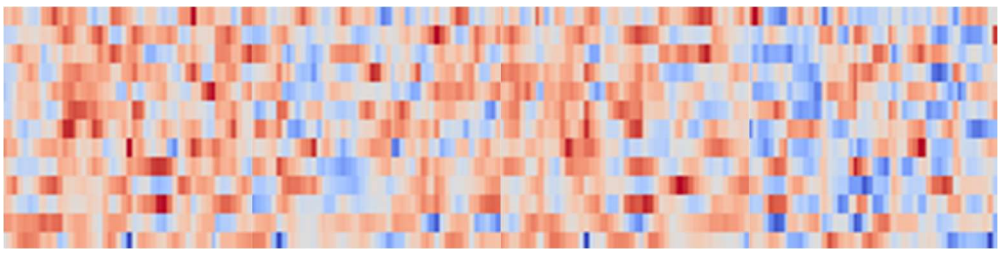
Chromagram



Mel spectrogram



MFCC rescaled



MFCC without rescaling

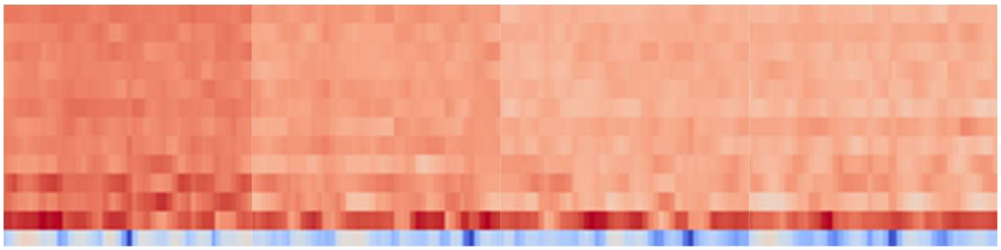


*Figure 4.20 Graphical representation of recordings used in the speech type detection procedure*

All 2D representations were tested with and without augmentation. Augmentation means using the alpha channel to store additional data:

- Gender of the speaker;
- F0 frequency;
- First two MFCCs.

These data are stored as information on a scale of 0-255 (like pixels on RGB color layers) on consecutive groups of pixels (roughly ¼ of the transparency layer for each of the above features).

Two different models were used; they are presented in Table 4.15. Initially, a third model with an additional dense layer was tested, but it increased the general complexity of the network (increasing the number of trainable parameters) and did not result in any improvement in overall accuracy.

*Table 4.15 Models used in speech type detection process*

**Model 1**

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_12 (Conv2D) | (None, 90, 93, 32) | 2080 |
| max_pooling2d_12 (MaxPooling | (None, 45, 46, 32) | 0 |
| dropout_20 (Dropout) | (None, 45, 46, 32) | 0 |
| conv2d_13 (Conv2D) | (None, 45, 46, 48) | 24624 |
| max_pooling2d_13 (MaxPooling | (None, 22, 23, 48) | 0 |

| Layer (type) | Output Shape | Param # |
|---|---|---|
| dropout_21 (Dropout) | (None, 22, 23, 48) | 0 |
| flatten_4 (Flatten) | (None, 24288) | 0 |
| dense_12 (Dense) | (None, 256) | 6217984 |
| dropout_22 (Dropout) | (None, 256) | 0 |
| dense_13 (Dense) | (None, 2) | 514 |

Total params: 6,245,202
Trainable params: 6,245,202
Non-trainable params: 0

**Model 2**

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_14 (Conv2D) | (None, 90, 93, 32) | 1184 |
| max_pooling2d_14 (MaxPooling | (None, 45, 46, 32) | 0 |
| dropout_23 (Dropout) | (None, 45, 46, 32) | 0 |
| conv2d_15 (Conv2D) | (None, 45, 46, 48) | 13872 |
| max_pooling2d_15 (MaxPooling | (None, 22, 23, 48) | 0 |
| dropout_24 (Dropout) | (None, 22, 23, 48) | 0 |
| conv2d_16 (Conv2D) | (None, 22, 23, 64) | 27712 |
| max_pooling2d_16 (MaxPooling | (None, 11, 11, 64) | 0 |
| dropout_25 (Dropout) | (None, 11, 11, 64) | 0 |
| flatten_5 (Flatten) | (None, 7744) | 0 |
| dense_14 (Dense) | (None, 256) | 1982720 |
| dropout_26 (Dropout) | (None, 256) | 0 |

```
_________________________________________________________________
dense_15 (Dense)             (None, 2)                 514
=================================================================
Total params: 2,026,002
Trainable params: 2,026,002
Non-trainable params: 0
```

Every model presented in Table 4.15 differ in several important features:

- Nuber of filters on the subsequent layers;
- Number of neurons in the last dense layer;
- Max pooling size;
- Dropout parameter value.

The configurations used are presented in Table 4.16.

*Table 4.16 Configuration of the learning process for speech type detection*

| ID | Model | Number of filters in conv. layers | Size of kernel | MaxPooling | Dropout after conv. layers | Dropout after dense layer | Number of neurons in the dense layer | Batch size | Number of epochs |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 16, 32 | 2 | 2 | 0.3 | 0.5 | 256 | 64 | 35 |
| 2 | 1 | 32, 48 | 2 | 2 | 0.3 | 0.5 | 256 | 64 | 35 |
| 3 | 1 | 32, 48 | 3 | 2 | 0.3 | 0.5 | 256 | 64 | 35 |
| 4 | 1 | 32, 48 | 3 | 2 | 0.3 | 0.5 | 256 | 32 | 35 |
| 5 | 1 | 32, 48 | 5 | 2 | 0.3 | 0.5 | 256 | 64 | 35 |
| 6 | 1 | 32, 48 | 5 | 2 | 0.3 | 0.5 | 256 | 32 | 35 |
| 11 | 2 | 32, 48, 64 | 3 | 2 | 0.3 | 0.5 | 256 | 64 | 35 |
| 12 | 2 | 32, 48, 64 | 5 | 3 | 0.3 | 0.5 | 256 | 64 | 35 |
| 13 | 2 | 32, 48, 64 | 5 | 3 | 0.3 | 0.5 | 256 | 32 | 35 |

The convolutional neural networks prepared using the configuration presented in Table 4.16 were trained using different graphical representations presented earlier (see Figure 4.20). The results of the experiments are presented in Table 4.17.

*Table 4.17 A summary of results of the speech type detection experiments*

| Type of graphical representation | Is the picture augmented or clean | Configuration ID | Accuracy of the testing set |
|---|---|---|---|
| mel spectrogram | augmented | 13 | 0.8671875 |
| spectrogram | clean | 12 | 0.8515625 |
| spectrogram | clean | 6 | 0.84375 |
| spectrogram | clean | 4 | 0.8359375 |
| spectrogram | augmented | 12 | 0.828125 |
| spectrogram | augmented | 13 | 0.828125 |
| spectrogram | augmented | 11 | 0.8125 |
| mel spectrogram | clean | 6 | 0.8125 |
| spectrogram | clean | 11 | 0.8046875 |
| spectrogram | augmented | 5 | 0.8046875 |
| mel spectrogram | clean | 11 | 0.8046875 |
| mel spectrogram | augmented | 6 | 0.8046875 |
| mel spectrogram | clean | 5 | 0.796875 |
| spectrogram | clean | 5 | 0.7890625 |
| mel spectrogram | clean | 12 | 0.7890625 |
| mel spectrogram | clean | 13 | 0.7890625 |
| mel spectrogram | augmented | 2 | 0.7890625 |

| | | | |
|---|---|---|---|
| mel spectrogram | augmented | 4 | 0.7890625 |
| spectrogram | clean | 2 | 0.78125 |
| spectrogram | clean | 3 | 0.78125 |
| spectrogram | augmented | 3 | 0.78125 |
| mel spectrogram | augmented | 11 | 0.78125 |
| spectrogram | clean | 13 | 0.765625 |
| spectrogram | augmented | 4 | 0.765625 |
| mel spectrogram | augmented | 12 | 0.765625 |
| mel spectrogram | clean | 4 | 0.7578125 |
| MFCC | clean | 4 | 0.7578125 |
| spectrogram | clean | 1 | 0.75 |
| spectrogram | augmented | 1 | 0.75 |
| mel spectrogram | clean | 2 | 0.75 |
| mel spectrogram | augmented | 3 | 0.75 |
| mel spectrogram | clean | 3 | 0.7421875 |
| chromagram | clean | 6 | 0.734375 |
| mel spectrogram | clean | 1 | 0.734375 |
| spectrogram | augmented | 2 | 0.703125 |
| chromagram | clean | 1 | 0.6953125 |
| spectrogram | augmented | 6 | 0.6796875 |
| mel spectrogram | augmented | 1 | 0.6796875 |
| chromagram | augmented | 12 | 0.671875 |

| | | | |
|---|---|---|---|
| chromagram | augmented | 13 | 0.671875 |
| MFCC | clean | 3 | 0.6640625 |
| MFCC | augmented | 2 | 0.6640625 |
| MFCC | clean | 2 | 0.65625 |
| chromagram | clean | 2 | 0.6328125 |
| chromagram | clean | 3 | 0.625 |
| chromagram | augmented | 11 | 0.625 |
| chromagram | clean | 11 | 0.6171875 |
| MFCC | clean | 6 | 0.6171875 |
| chromagram | clean | 12 | 0.609375 |
| chromagram | clean | 13 | 0.6015625 |
| chromagram | augmented | 4 | 0.6015625 |
| MFCC | clean | 12 | 0.5859375 |
| MFCC | augmented | 1 | 0.5859375 |
| MFCC | augmented | 5 | 0.5859375 |
| chromagram | augmented | 3 | 0.578125 |
| mel spectrogram | augmented | 5 | 0.578125 |
| MFCC | augmented | 12 | 0.5703125 |
| chromagram | augmented | 5 | 0.5546875 |
| MFCC rescaled | augmented | 12 | 0.546875 |
| MFCC | clean | 13 | 0.546875 |
| MFCC | augmented | 6 | 0.546875 |

| | | | |
|---|---|---|---|
| MFCC | augmented | 11 | 0.5390625 |
| chromagram | augmented | 1 | 0.53125 |
| MFCC | clean | 5 | 0.53125 |
| chromagram | augmented | 2 | 0.5234375 |
| chromagram | augmented | 6 | 0.515625 |
| MFCC rescaled | clean | 2 | 0.515625 |
| MFCC rescaled | clean | 6 | 0.515625 |
| MFCC rescaled | clean | 11 | 0.515625 |
| MFCC rescaled | clean | 12 | 0.515625 |
| MFCC rescaled | clean | 13 | 0.515625 |
| MFCC rescaled | augmented | 1 | 0.515625 |
| MFCC | clean | 1 | 0.515625 |
| MFCC | augmented | 3 | 0.515625 |
| MFCC rescaled | augmented | 2 | 0.5078125 |
| MFCC rescaled | clean | 1 | 0.5 |
| MFCC | augmented | 13 | 0.5 |
| MFCC rescaled | clean | 4 | 0.484375 |
| MFCC rescaled | clean | 5 | 0.484375 |
| MFCC rescaled | augmented | 5 | 0.484375 |
| MFCC rescaled | augmented | 6 | 0.484375 |
| MFCC | augmented | 4 | 0.4765625 |
| chromagram | clean | 5 | 0.46875 |

| | | | |
|---|---|---|---|
| chromagram | clean | 4 | 0.4609375 |
| MFCC rescaled | clean | 3 | 0.4609375 |
| MFCC rescaled | augmented | 4 | 0.4609375 |
| MFCC rescaled | augmented | 3 | 0.453125 |
| MFCC | clean | 11 | 0.453125 |
| MFCC rescaled | augmented | 11 | 0.4296875 |
| MFCC rescaled | augmented | 13 | 0.4140625 |

Table 4.17 shows that the mel spectrogram with augmentation is the best candidate for further processing. In contrast, the representations related to the MFCCs gave unsatisfactory results. This does not mean that the value of the features does not convey any information in this context − conversely − the augmented 2D representations use the first two MFCCs.

**Lombard speech detection using mel spectrogram and extended number of graphical samples, with augmentation (experiment 8)**

In the previous experiments, the number of training items was equal to the number of recordings (640) with regard to the fact that these samples were divided into training, validation, and test sets. Effectively training was performed using around 400 recordings (400 graphics), which resulted in lower recognition performance than was expected.

This time the data have been prepared in another way:

- Every sound file has been read using 22050 Hz sampling frequency
- Average F0 was calculated for the whole file
- Average MFCCs have been calculated (second and third coefficient)
- Every file has been divided into windows of length around 1/3 of the sampling frequency (around 7000 samples) and the hop length of 2000 samples (which means that the windows were overlapping).

For each fragment, the mel spectrogram was calculated, and a graphical representation was generated if at least 90% of the fragment carries energy information (to avoid training the network on segments where the majority of them contain silence).

Sample mel spectrograms are presented in Figure 4.21.

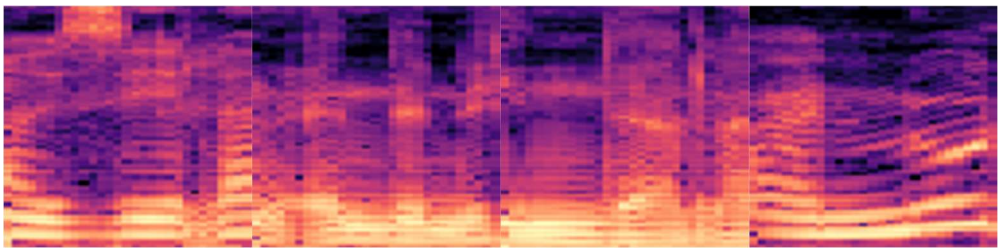


*Figure 4.21 Sample mel spectrograms used in experiment 8*

This way, 4933 mel spectrograms (each of about 7000 time-domain samples) were obtained based on 640 record files.

The model on which the network was trained (during training, the data was augmented, i.e., using a transparency layer to store information about gender, F0, and MFCCs) is presented in Table 4.18.

*Table 4.18 CNN model used in experiment 8*

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_20 (Conv2D) | (None, 90, 93, 32) | 3232 |
| max_pooling2d_20 (MaxPooling | (None, 30, 31, 32) | 0 |
| dropout_32 (Dropout) | (None, 30, 31, 32) | 0 |
| conv2d_21 (Conv2D) | (None, 30, 31, 48) | 38448 |
| max_pooling2d_21 (MaxPooling | (None, 10, 10, 48) | 0 |
| dropout_33 (Dropout) | (None, 10, 10, 48) | 0 |
| conv2d_22 (Conv2D) | (None, 10, 10, 64) | 76864 |
| max_pooling2d_22 (MaxPooling | (None, 3, 3, 64) | 0 |
| dropout_34 (Dropout) | (None, 3, 3, 64) | 0 |

```
flatten_7 (Flatten)          (None, 576)               0
_________________________________________________________________
dense_19 (Dense)             (None, 512)               295424
_________________________________________________________________
dropout_35 (Dropout)         (None, 512)               0
_________________________________________________________________
dense_20 (Dense)             (None, 256)               131328
_________________________________________________________________
dropout_36 (Dropout)         (None, 256)               0
_________________________________________________________________
dense_21 (Dense)             (None, 2)                 514
=================================================================
Total params: 545,810
Trainable params: 545,810
Non-trainable params: 0
_________________________________________________________________
```

It can be clearly seen that the network has relatively few trained parameters due to the reasonably large max-pooling (3). The number of epochs is 60, and the batch size is 32. Accuracy on the test set, however, is very high: 98.3%, and the loss is at 0.05.

Sample classifications are presented in Figure 4.22.

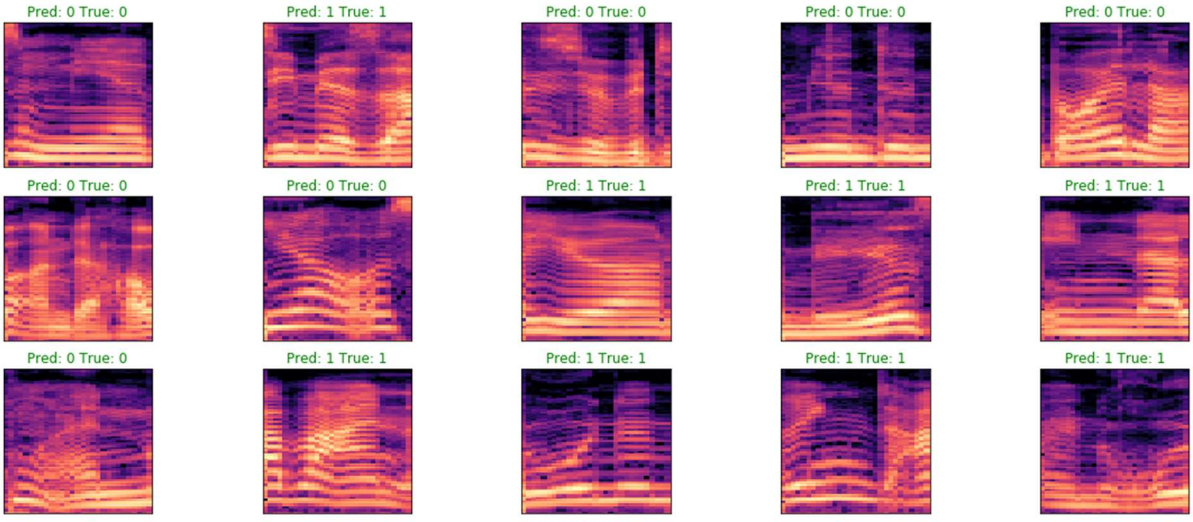



*Figure 4.22 Sample classifications for experiment 8*

**Experiment 9 − repeated experiment 8, without augmentation**

To compare the importance of augmentation, experiment 8 was repeated with an identical model and hyperparameter values, but augmentation was removed from the training process.

The effect is much worse - accuracy on the test set is 90%, and loss is 0.23. Sample classification results are presented in Figure 4.23.

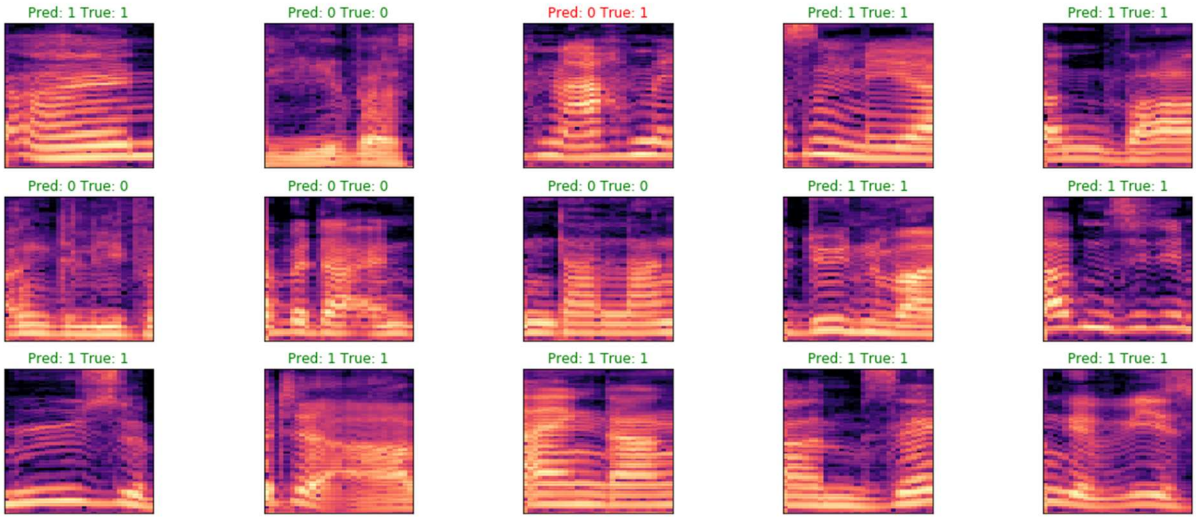



*Figure 4.23 Sample classifications for experiment 9*

## 4.2.4 Implementation of the detection process

**Implementation process**

The process of implementing the detection method has been divided into three stages:

- Preparing mel spectrograms according to the method described above. All the pictures taken are indexed in a single file, containing - apart from the access path to the picture - also information about the speaker's gender, the presence of noise during the recording, the F0 frequency, and two MFCCs.
- Performing convolutional network training with the use of created 2D representations. Training lasts 60 epochs, and a model was saved that provides the highest accuracy on the validation set.
- Recognition testing on all recordings. Recognition, in this case, concerns a single fragment of the tested recording, prepared in the same way as the pictures for training.

The entire procedure was repeated for German and Polish recordings.

**Averaging the recognition results**

It should be noted that the process of recognizing the type of speech will be dynamic as you speak and does not involve a longer fragment of speech but only a short piece of it (e.g., 0.25 seconds). It is often a fragment that does not carry too much energy (e.g., a moment of silence) or an ambiguous fragment (e.g., one in which a large part of the time contains silence and non-energetic phonemes). Therefore, it should be taken into account that the recognition of the speech type will not be the same for the entire course of the tested speech signal. There is a high probability that the nature of such speech changes as the sentence is uttered.

To avoid misclassification due to a temporary change in the nature of the utterance, averaging the results is an important element of the recognition process. In the case of real-time recognition, the process memory can be used, e.g., averaging the results for the last dozen or so windows. For the purposes of this work, the results for the entire recording subjected to tests were averaged because the recognition result changes during the recording due to silence or unvoiced fragments.

Sample prediction (detection) diagrams are presented in Figure 4.24.

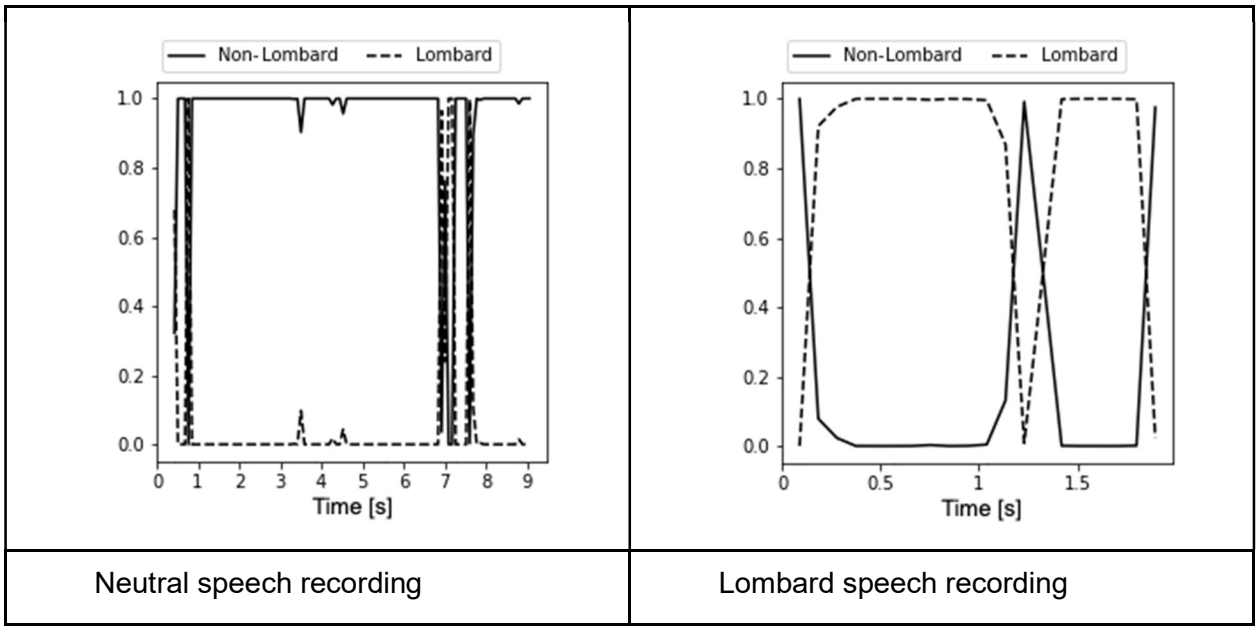



*Figure 4.24 Sample detection diagrams for neutral and Lombard speech*

On the horizontal axis, there is the number of the classified frame; on the vertical axis, the probability of the fact that a given window is a fragment of Lombard (dashed line) or neutral (solid line) speech.

The left side of Fig. 4.24 shows a typical recording of non-Lombard speech - most of the frames of the recording have been classified as neutral recording. The right side of Fig. 4.24 presents a recording of the Lombard speech - the fluctuations are much greater, but the advantage of frames classified as Lombard speech is visible.

From the point of view of the classification of the recording as a whole, the average recognition value is essential. It was performed as follows:

Let us assume that the threshold for classifying a given frame as Lombard is 0.5 – i.e., if the neural network returns the vector [*N, L*] for each frame, it actually returns two probabilities: with probability *N,* the given frame is neutral speech, with probability *L* – Lombard speech. If *N*> L, the frame is non-Lombard; otherwise – Lombard-like. The result of this comparison is the value of NOISE – for NOISE = 0 – neutral speech, for NOISE = 1 – Lombard speech.

The result of the above comparison is the vector X, the length of which depends on the number of classified frames. It should be mentioned that all frames are classified whether or not they contain speech.

The classification of the entire recording will result in the average of *A = AVG (X).* The obtained value is then compared with an empirically defined level Y. This level is called the cutoff level, and the classification result is defined according to it.

A sample chart of the averaged detection results is shown in Figure 4.25.

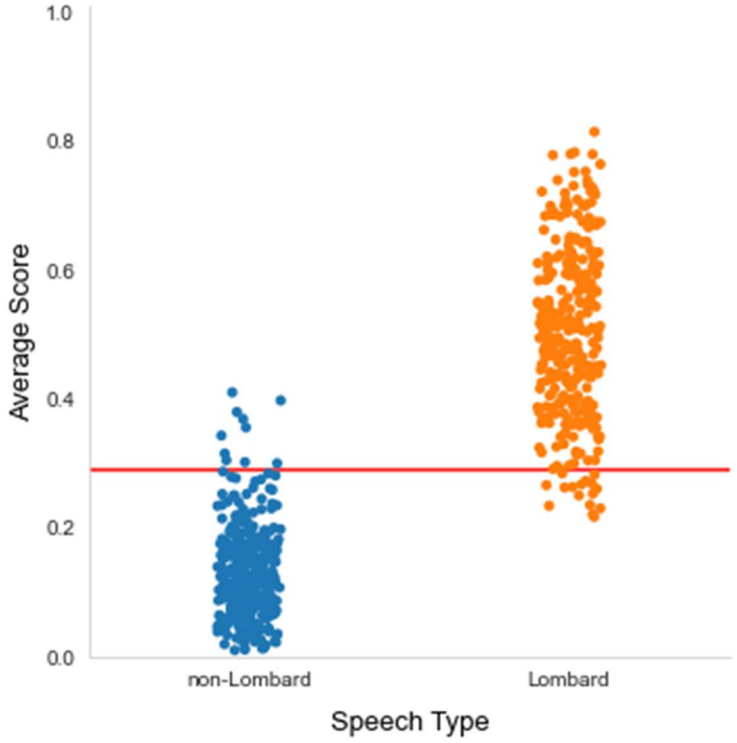



*Figure 4.25 Averaged scores of the Lombard-speech prediction*

It can be seen from the graph that a good cutoff (Y) is around 0.3. Of course, this means that some recordings will be incorrectly classified as neutral speech (false negative) or as Lombard speech (false positive). Minimizing both indicators is one of the goals of determining the correct cutoff level.

### 4.2.5 Evaluation results

**Procedure**

Before the approach to testing the recognition engine, the following components were prepared in advance:

- Model of a trained network with weights retained for optimal network accuracy for the validation set.
- A set of data on recording frames used in network training. The data includes information on F0 and the first two MFCCs for the entire recordings. This allows the data to be brought into the range of 0-255, which is typical for pictures.
- Recordings for recognition testing.

Three convolutional neural network models were trained for testing. For each of them, the generation parameters are as follows:

- Fmax = 8000 Hz – maximum frequency of the mel filter bank.
- Divider = 3 – a divisor that affects the length of the frame used to generate it. a picture for the purpose of training the neural network (the frame is 22050 Hz / sample divider)
- Step = 2000 – shift step between frames.

The following network models were created:

- Networks trained on German recordings:
    - A network trained on the recordings of all eight speakers and tested on the same recordings (model G1).
    - A network trained on the recordings of six speakers and tested on the other two (G2 model).
- Networks trained on Polish recordings:
    - A network trained on the recordings of all four Polish speakers and tested on the same recordings (P1 model).

**Results of the training and detection**

*Table 4.19 Training and detection results*

| Model | G1 | G2 | P1 |
|---|---|---|---|
| Number of samples used for training | 3156 | 2334 | 816 |
| Number of samples used for validation | 790 | 584 | 205 |
| Accuracy on validation set | 0.9899 | 0.9880 | 0.9902 |
| Loss on validation set | 0.0370 | 0.0434 | 0.0432 |
| Cutoff level | 0.29 | 0.20 | 0.68 |
| Recognition accuracy | | | |
| Accuracy | 0.9594 | 0.9406 | 0.9667 |
| True positives | 304 | 124 | 57 |
| True negatives | 310 | 240 | 59 |
| False positives | 10 | 6 | 1 |
| False negatives | 16 | 17 | 3 |
| Precision | 0.9681 | 0.9538 | 0.9828 |
| Recall | 0.9500 | 0.8794 | 0.9500 |

**Visualizations of the model separation**

Model values separation can be visualized using scatter plots, presenting all recognized (detected) speech types with their average detection score. Separation plots are presented in Figures 4.26 – 4.28.

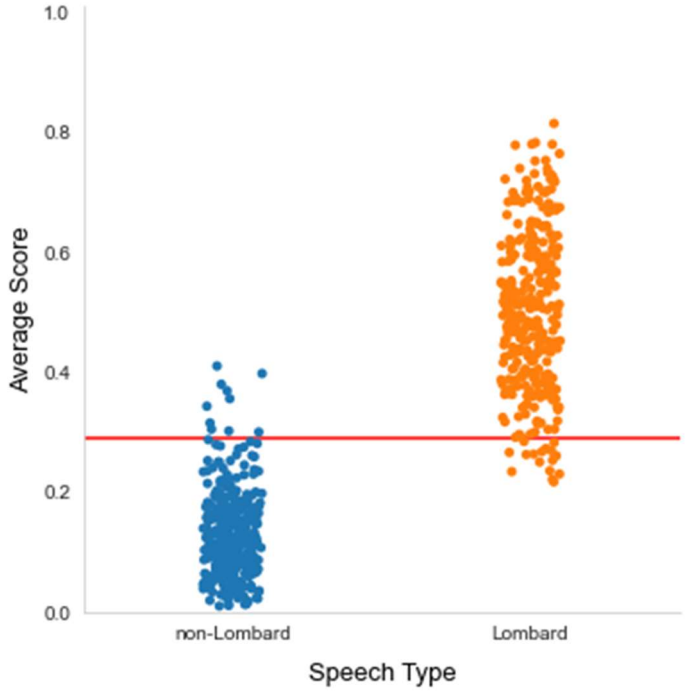



*Figure 4.26 G1 model with the cutoff level of 0.29*

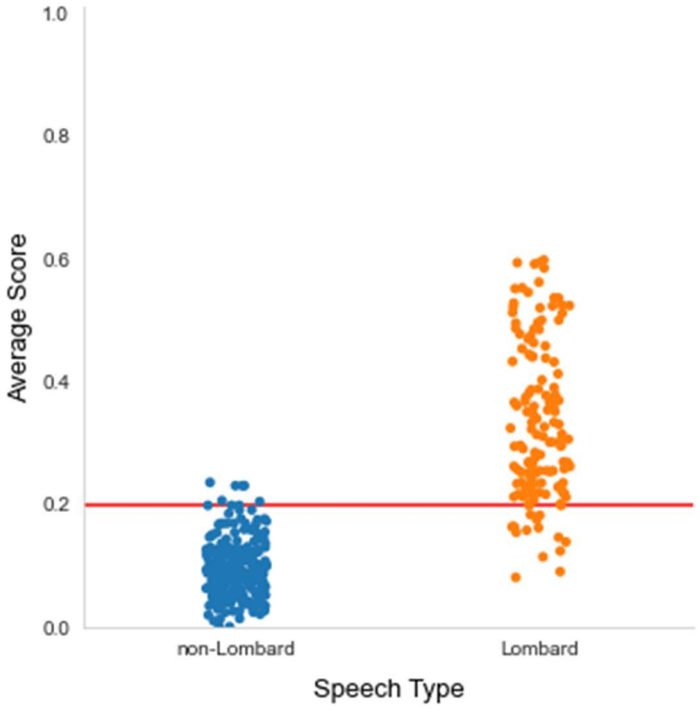



*Figure 4.27 G2 model with the cutoff level of 0.2*

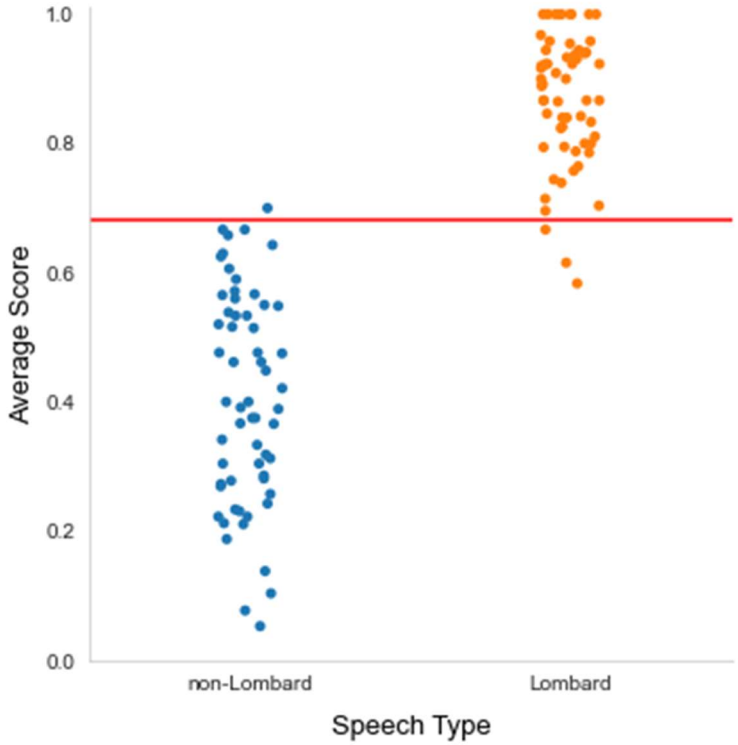



*Figure 4.28 P1 model with the cutoff level of 0.68*

As can be seen from the above charts, there is a clear separation between Lombard and neutral recordings. It can successfully be used to implement the decision component.

**Sample visualizations of the speech signal**

Visualizations presented in Figures 4.29 to 4.32 consist of three pictures:

- Speech signal (recording visualizations);
- Mel spectrogram;
- Prediction diagram.

Figures 4.29 and 4.30 present the visualizations of utterances recorded by a Polish speaker. Figures 4.31 and 4.32 present the visualizations of German utterances. There is a difference in sound level for both languages, but this was caused by the recording setup used in both corpora (Czyzewski *et al.*, 2017; Soloducha *et al.*, 2016).

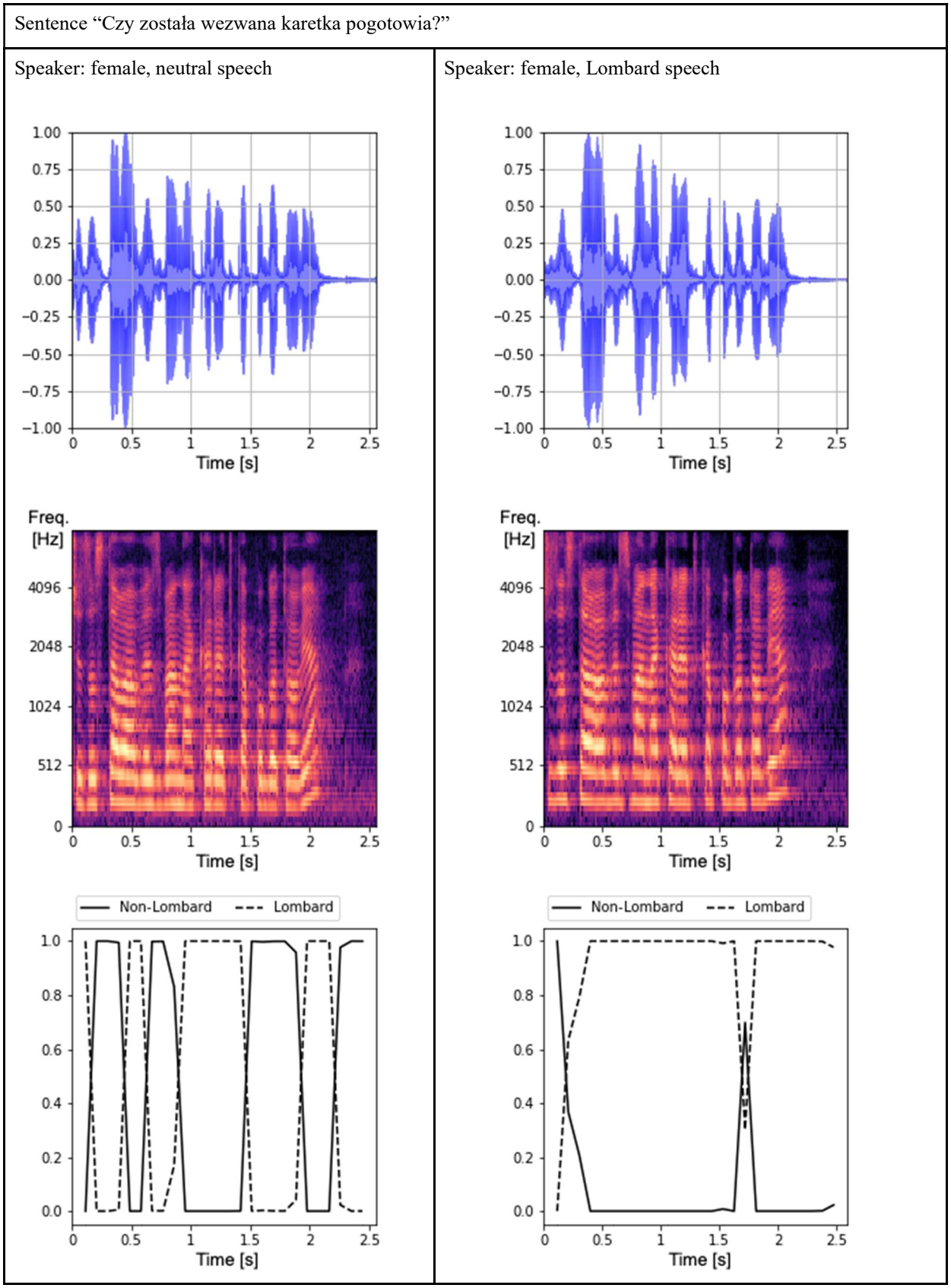



*Figure 4.29 Polish recording number 1*

| Sentence “Czy została wezwana karetka pogotowia?” | |
|---|---|
| Speaker: male, neutral speech | Speaker: male, Lombard speech |

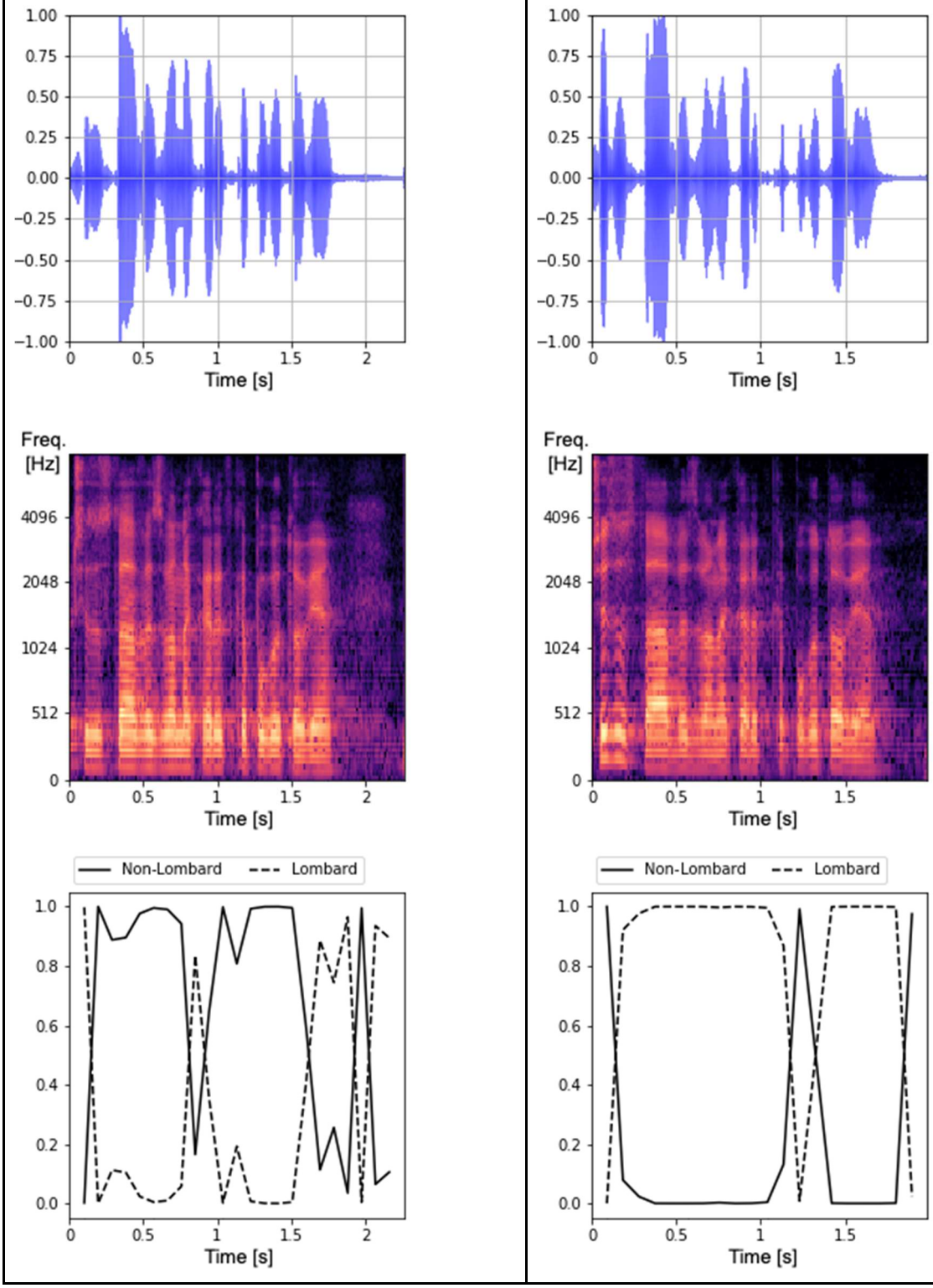



*Figure 4.30 Polish recording number 2*

| Sentence “Ich bin in Brüssel angekommen, aber mein Gepäck in Jena. Ich bin auf mein Gepäck dringend angewiesen, weil ich Diabetiker bin. Informieren sie bitte die zuständige Stelle.” | |
|---|---|
| Speaker: female, neutral speech | Speaker: female, Lombard speech |

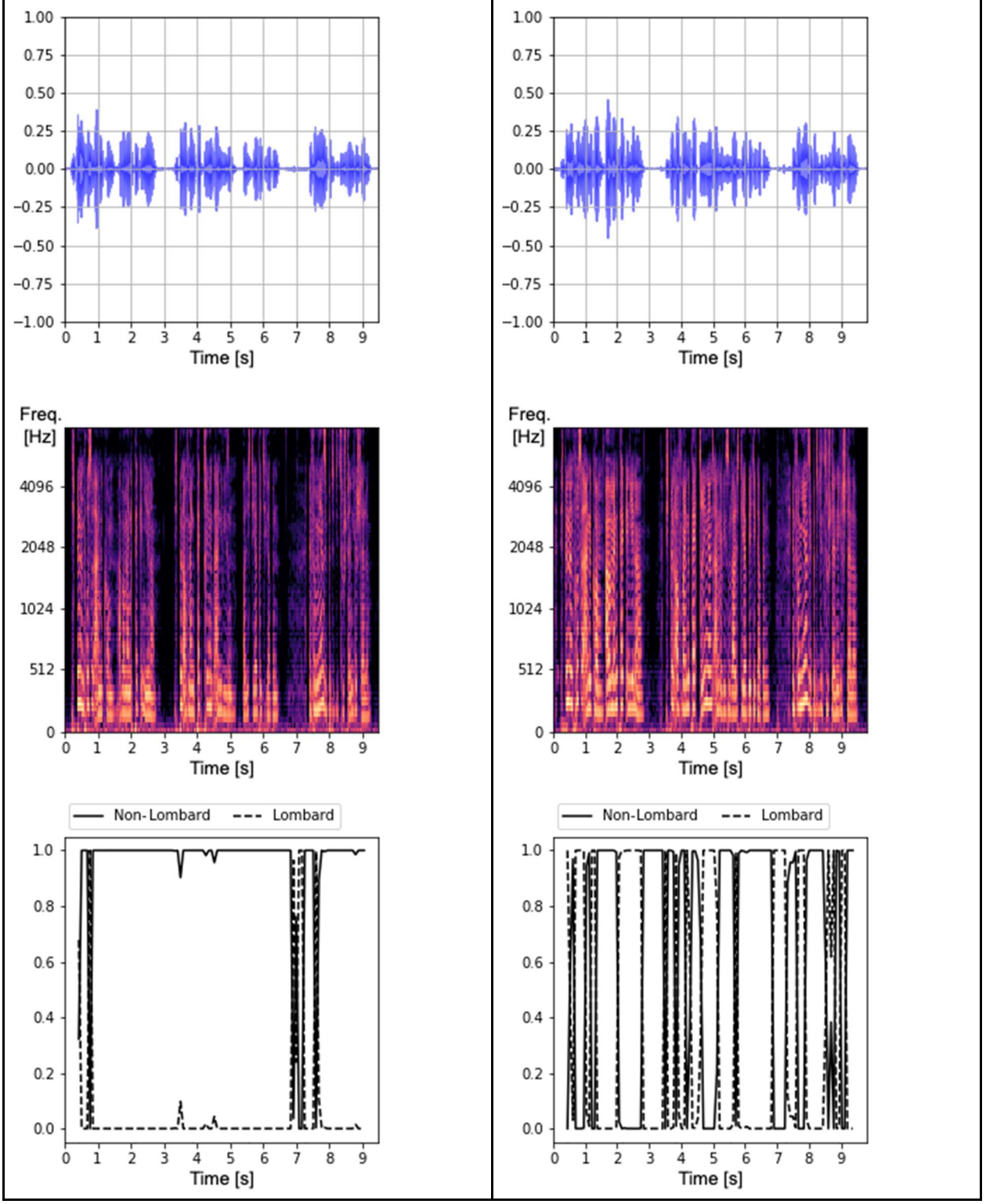



*Figure 4.31 German recording number 1*

| Zdanie “Ich bin in Brüssel angekommen, aber mein Gepäck in Jena. Ich bin auf mein Gepäck dringend angewiesen, weil ich Diabetiker bin. Informieren sie bitte die zuständige Stelle.” | |
|---|---|
| Speaker: male, neutral speech | Speaker: male, Lombard speech |

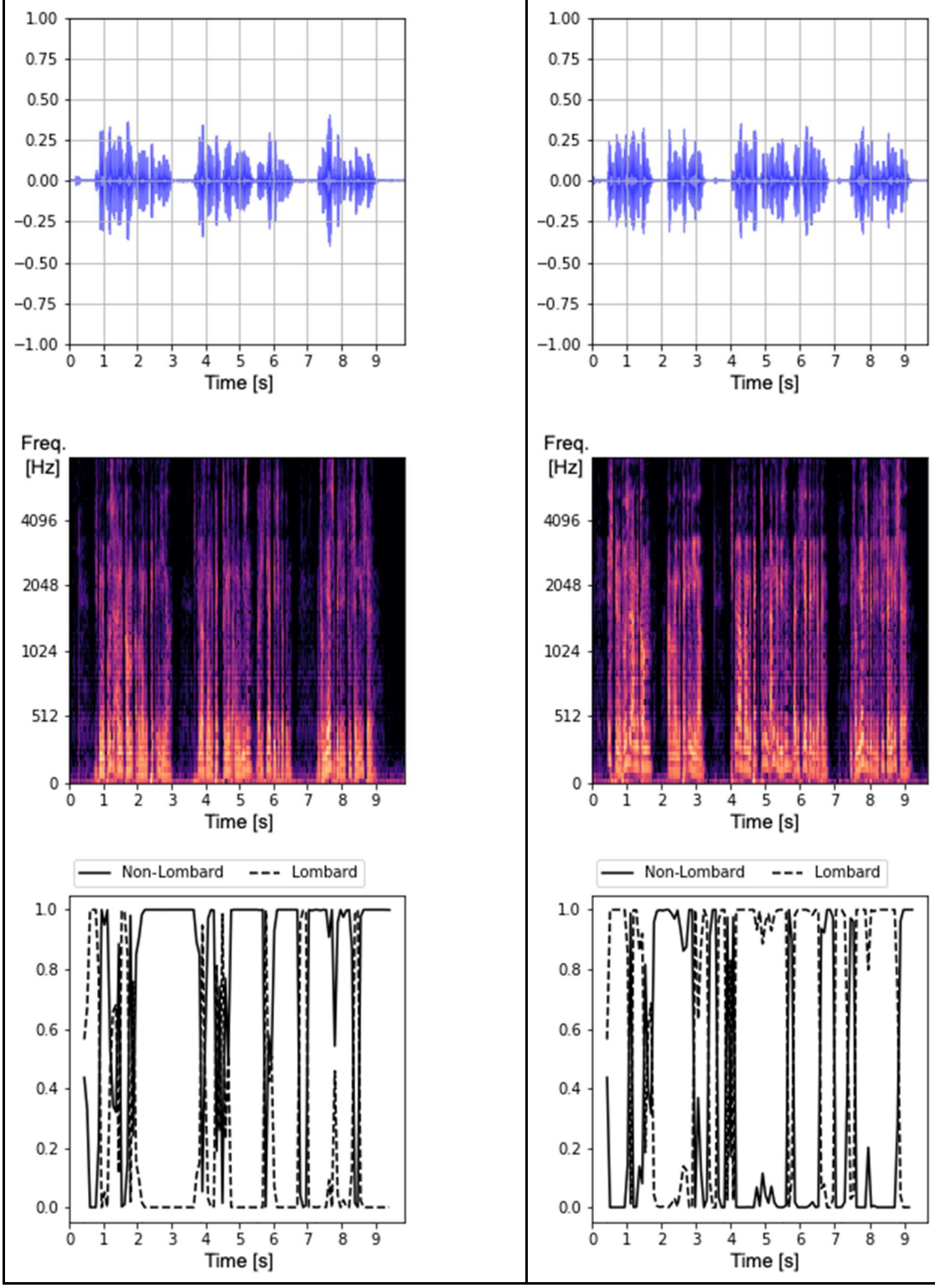



*Figure 4.32 German recording number 2*

### 4.2.6 Conclusions

The deep learning method using convolutional networks is a very convenient approach to the problem of Lombard speech detection. Interestingly, the convolutional network provides such good results that it is difficult to see, even with an expert eye, the differences, i.e., Lombard and neutral (non-Lombard) speech, detected by the network. This way, thesis no. 1 has been proved.

The use of a trained network is limited to a specific language or setup of recording devices. The characteristics of the recorded signal affect the learning ability of the network. Hence, for the correct operation of such a component that recognizes Lombard speech, it is necessary to train the network on a specific set of voice recording devices. It is also impossible to use a network trained on German recordings to recognize Polish recordings.

After training the network for a specific application, a cutoff point should be established to increase recognition accuracy. In each case, this threshold may be different.

The generated mel spectrograms in the form of pictures are used in the recognition process. These pictures are not generated in the form of physical files but as visualizations are saved in the memory as a byte array. From the point of view of the convolutional network, this is not necessary - the convolutional network treats data in the same way, whether it is an image or an ordinary data tensor. Therefore, the use of visualization is not necessary here, and resignation from it will accelerate the learning and recognition process.

Information about the gender of the speaker was used to identify the Lombard speech. It should be underlined that such information is not available a priori in real-time systems. It is, therefore, necessary to add a component that recognizes the speaker's gender - preferably based on a separate classification mechanism (based on F0 or mel spectrograms). Of course, the simplicity of the system may be more important than its accuracy – in such a case – the gender information might be ignored (this is the case described in this dissertation).

In the experiments, Lombard speech recognition was performed for the entire recording. If such a component were to be used in real-time systems, recognition must also be performed in real-time, with some delay from the start (i.e., before the first recognition is made, there must be a short moment without a decision to collect part of the speech signal already). It also means the necessity to develop a practical averaging algorithm and memory of previous diagnoses.

In real voice systems, recognition must consider the silence between utterances and ignore these passages (Makowski and Hossa, 2020; Wang and Nishizaki, 2022). Therefore, there is a need to use a separate VAD system or extend the developed system to the three-valued classification. Then, in addition to recognizing Lombard or neutral speech, the system could be able to recognize silence and be a self-sufficient component of the VAD.

## 4.3 Noise profiling

This Chapter aims to prepare the machine-based model to recognize the noise type and correctly classify it near real-time. Based on noise classification, it will be possible to modify the speech signal appropriately to increase the probability of improving its quality and intelligibility. The experiments are conducted with a new perspective, focusing not on assigning a disturbance to a class but rather on investigating the stability of this assignment. This constitutes a new measure of the quality of profiling that is time-dependent. This research area requires a thorough analysis of speech and noise elements based on a microscopic scale. Therefore, the large-scale deep learning analysis is left outside in this research, disregarding that noise robustness is well-served by deep learning methods (Watanabe *et al.*, 2017). However, state-of-the-art baseline algorithms comprising the extraction of features and machine learning such as Naïve Bayes, linear SVM (Support Vector Machine), SVM with the polynomial kernel, Gaussian process classifiers, decision tree, random forest, MLP (Multilayer Perceptron), AdaBoost classifier, and finally Quadratic Discriminant Analysis are used. It is worth noting that the methodology based on feature extraction and baseline classifiers shows its superiority in many speech signal processing tasks (Bhavan *et al.*, 2019; Tuncer *et al.*, 2021).

Figure 4.33 shows the current Chapter topic coverage.

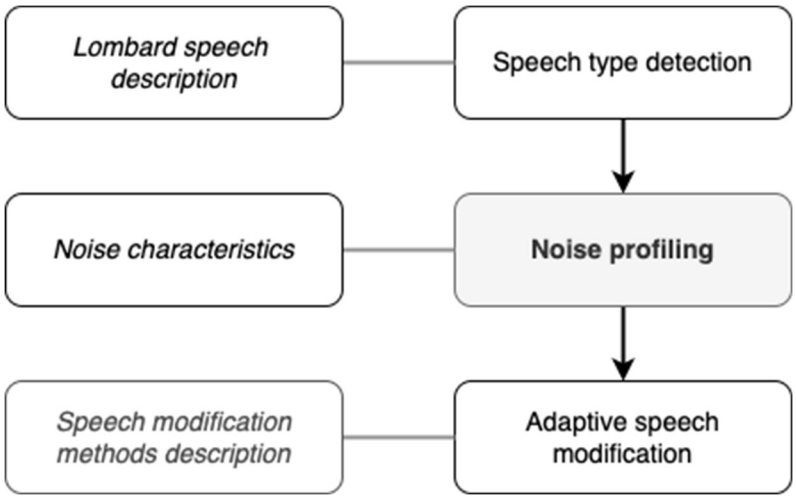



*Figure 4.33 Structure of the dissertation with current Chapter topics highlighted*

In this dissertation work, it is justified why the process of improving speech quality and intelligibility should be adaptive and specific modifications may depend on the noise characteristics and reinforced by them. Therefore, a stable noise profiling method is needed.

Possible speech modifications must fit the disruption to provide the best results in terms of potential loss in intelligibility because of the noise presence. It is because every disturbance has different characteristics and impacts speech differently. In the experiment, noise profiling is based on signal frequency analysis. However, it is not crucial to recognize a given type of noise (for instance, recognizing that a given noise signal is a babble speech or airport noise), but – as already mentioned – it is critical to have this process repetitive and stable. Also, noise signals with similar frequency characteristics should always be analogously classified to ensure that the speech signal modification is appropriate and durable.

### 4.3.1 Material and methods

In the learning process, the Aurora noise dataset was used (Ellis, 2002). The noise signals contained in the Aurora dataset are as follows: airport, babble speech, car noise, exhibition, restaurant, street noise, subway, and train. In addition, pink noise was generated as this noise type was not present in the Aurora database). The following frequency characteristics were chosen and extracted to classify noise types (Dubnov, 2004; Klapuri and Davy, 2006; McFee *et al.*, 2015), i.e., spectral centroid, spectral bandwidth, spectral flatness.

The most important factor in evaluating the usefulness of the given feature is the separation of the calculated values in the context of the considered noise type. Three frequency characteristics, described in Chapter 2.3.2, calculated in real-time, were considered to increase the separation of different types of noise. What is more, for each of the characteristics, the following short-term statistical parameters are calculated: maximum value, minimum value, average value, amplitude, standard deviation, variance, and median. The given statistical values should provide great noise parameters separation. The frequency characteristics are calculated from the Fourier spectrum computed with a Hamming window of 2048 samples (25% overlap).

### 4.3.2 Noise analyses

As presented in Chapter 2.3.2, the following parameters were analyzed for noise profiling: spectral bandwidth, spectral flatness, and spectral centroid. Figures 4.34 to 4.36 present spectral characteristics for every type of tested noise signal.

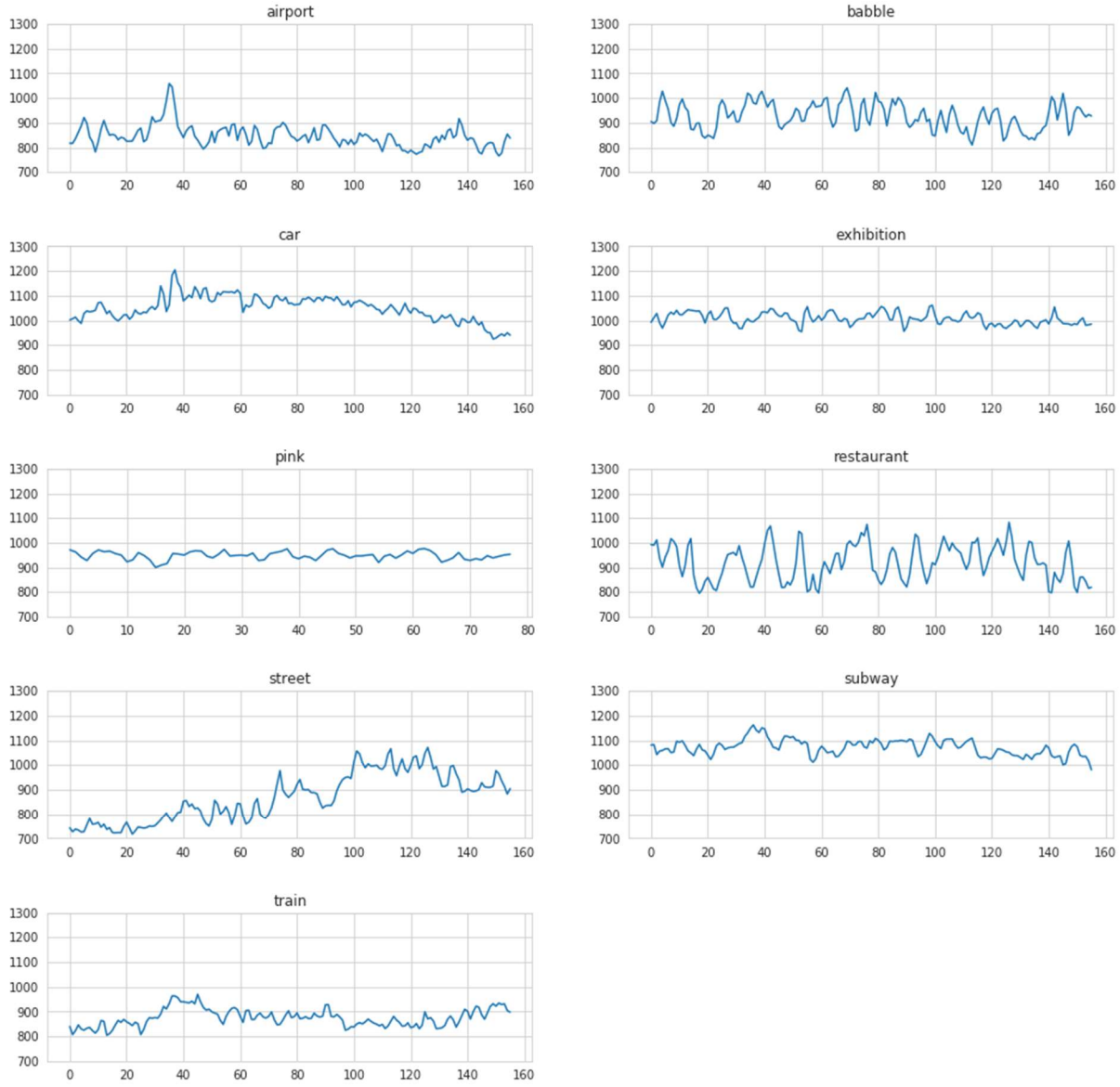



*Figure 4.34 Spectral bandwidth charts of noise recordings*


*Figure 4.35 Spectral flatness of noise recordings*

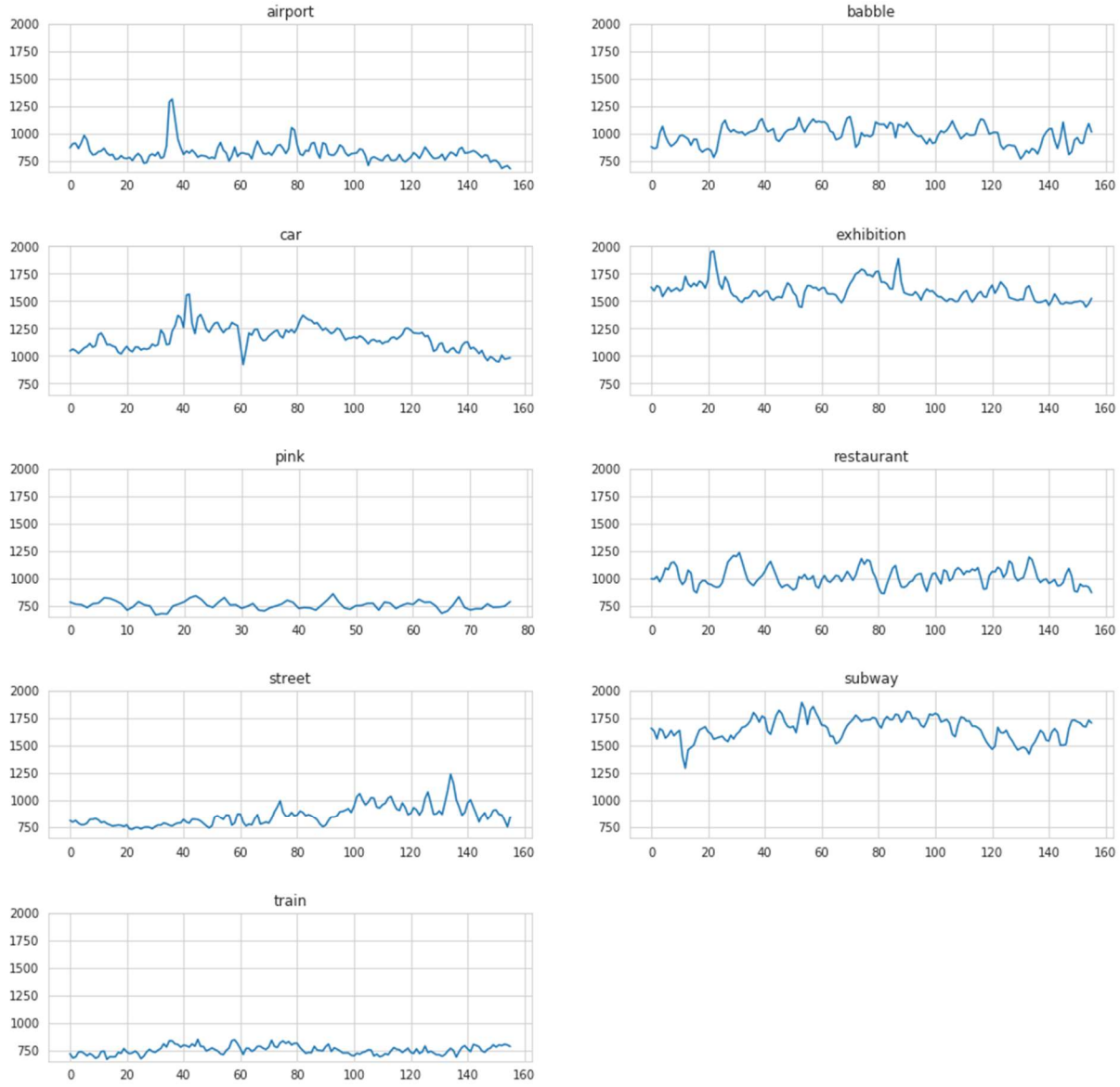



*Figure 4.36 Spectral centroid of noise recordings*

### 4.3.3 Noise type recognition model

Based on the previously described frequency characteristics, the recognition model was built, using the models presented in Chapter 2.3.1. For that purpose, several baseline algorithms were employed, i.e., Naïve Bayes (Barber, 2011; Zhang, 2004), linear SVM (Cortes and Vapnik, 1995; Platt, 1999), SVM with the polynomial kernel (Wu *et al.*, 2004), Gaussian process classifiers (Byrd *et al.*, 1995; Rasmussen and Williams, 2006; Zhu *et al.*, 1997), decision tree (Kamiński *et al.*, 2018), random forest (Ho, 1995), MLP (Multilayer Perceptron) (Pedregosa *et al.*, 2011), AdaBoost classifier (Rojas, 2009), and Quadratic Discriminant Analysis (Cortes *et al.*, 2004; Fernández-Delgado *et al.*, 2014; James, n.d.) that arose from different families and areas of

knowledge (Fernández-Delgado *et al.*, 2014). Every recording containing noise was processed in the following way:

- each frame is 2 seconds in length - to retrieve the statistical features for the training process,
- a 2-second window is moved by 0.1 seconds in each analysis step.

The classification models built use relatively long recording fragments because the measured parameter values change in time to a great extent.

### 4.3.4 Comparison of the classifier results

The classification results are provided in the form of overall accuracy and a confusion matrix, allowing for an easy interpretation of the results. Table 4.20 shows the comparison of the above-described classification models. Also, other metrics such as P – precision, R – recall, F1 – F1 score, S – support are included. The best accuracy and ROC AUC - area under the receiver operating characteristic – are highlighted in bold.

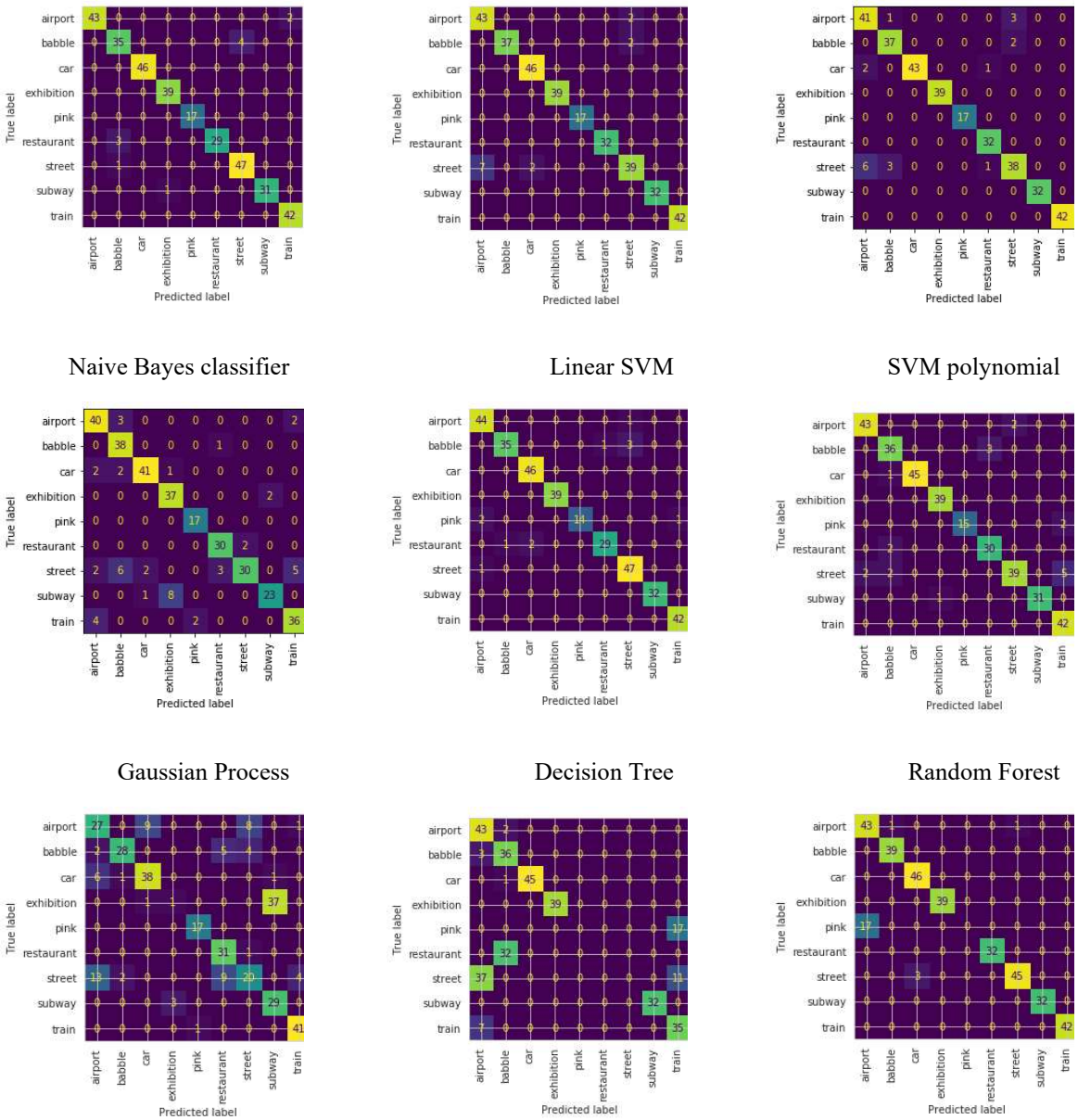
*Table 4.20 Results of the classification using different classification models. P – precision, R – recall, F1 – F1 score, S – support*

| | Naïve Bayes | | | | Linear SVM | | | | **SVM polynomial** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | **96.76%** | | | | 96.17% | | | | 94.41% | | | |
| **ROC AUC** | **0.99** | | | | **0.99** | | | | **0.99** | | | |
| **Noise distortions** | **P** | **R** | **F1** | **S** | **P** | **R** | **F1** | **S** | **P** | **R** | **F1** | **S** |
| Airport | 1.00 | 0.96 | 0.98 | 45 | 0.86 | 0.96 | 0.91 | 45 | 0.84 | 0.91 | 0.87 | 45 |
| Babble speech | 0.90 | 0.90 | 0.90 | 39 | 1.00 | 0.95 | 0.97 | 39 | 0.90 | 0.95 | 0.93 | 39 |
| Car | 1.00 | 1.00 | 1.00 | 46 | 0.96 | 1.00 | 0.98 | 46 | 1.00 | 0.93 | 0.97 | 46 |
| Exhibition | 0.98 | 1.00 | 0.99 | 39 | 1.00 | 1.00 | 1.00 | 39 | 1.00 | 1.00 | 1.00 | 39 |
| Pink noise | 1.00 | 1.00 | 1.00 | 17 | 1.00 | 1.00 | 1.00 | 17 | 1.00 | 1.00 | 1.00 | 17 |
| Restaurant | 1.00 | 0.91 | 0.95 | 32 | 1.00 | 1.00 | 1.00 | 32 | 0.94 | 1.00 | 0.97 | 32 |
| Street noise | 0.92 | 0.98 | 0.95 | 48 | 0.91 | 0.81 | 0.86 | 48 | 0.88 | 0.79 | 0.84 | 48 |
| Subway | 1.00 | 0.97 | 0.98 | 32 | 1.00 | 1.00 | 1.00 | 32 | 1.00 | 1.00 | 1.00 | 32 |
| Train | 0.95 | 1.00 | 0.98 | 42 | 1.00 | 1.00 | 1.00 | 42 | 1.00 | 1.00 | 1.00 | 42 |

| | Gaussian process | | | | Decision Tree | | | | Random Forest | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | 85.88% | | | | 96.47% | | | | 94.11% | | | |
| **ROC AUC** | 0.98 | | | | 0.98 | | | | **0.99** | | | |
| **Noise distortions** | **P** | **R** | **F1** | **S** | **P** | **R** | **F1** | **S** | **P** | **R** | **F1** | **S** |
| Airport | 0.83 | 0.89 | 0.86 | 45 | 0.94 | 0.98 | 0.96 | 45 | 0.98 | 0.98 | 0.98 | 45 |
| Babble speech | 0.78 | 0.97 | 0.86 | 39 | 0.97 | 0.77 | 0.86 | 39 | 0.87 | 1.00 | 0.93 | 39 |
| Car | 0.93 | 0.89 | 0.91 | 46 | 1.00 | 1.00 | 1.00 | 46 | 0.98 | 1.00 | 0.99 | 46 |
| Exhibition | 0.80 | 0.95 | 0.87 | 39 | 0.98 | 1.00 | 0.99 | 39 | 0.98 | 1.00 | 0.99 | 39 |
| Pink noise | 0.89 | 1.00 | 0.94 | 17 | 1.00 | 0.88 | 0.94 | 17 | 1.00 | 0.94 | 0.97 | 17 |
| Restaurant | 0.88 | 0.94 | 0.91 | 32 | 0.84 | 0.97 | 0.90 | 32 | 0.84 | 0.97 | 0.90 | 32 |
| Street noise | 0.94 | 0.63 | 0.75 | 48 | 0.92 | 0.98 | 0.95 | 48 | 1.00 | 0.58 | 0.74 | 48 |
| Subway | 0.92 | 0.72 | 0.81 | 32 | 1.00 | 0.97 | 0.98 | 32 | 1.00 | 0.97 | 0.98 | 32 |
| Train | 0.84 | 0.86 | 0.85 | 42 | 1.00 | 1.00 | 1.00 | 42 | 0.82 | 1.00 | 0.90 | 42 |

| | MLP Classifier | | | | AdaBoost Classifier | | | | QDA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | 68.23% | | | | 67.64% | | | | 93.52% | | | |
| **ROC AUC** | 0.95 | | | | 0.95 | | | | 0.94 | | | |
| **Noise Distortions** | **P** | **R** | **F1** | **S** | **P** | **R** | **F1** | **S** | **P** | **R** | **F1** | **S** |
| Airport | 0.75 | 0.40 | 0.52 | 45 | 0.48 | 0.96 | 0.64 | 45 | 0.72 | 0.96 | 0.82 | 45 |
| Babble speech | 0.74 | 0.74 | 0.74 | 39 | 0.51 | 0.92 | 0.65 | 39 | 0.98 | 1.00 | 0.99 | 39 |
| Car | 0.85 | 0.85 | 0.85 | 46 | 1.00 | 0.98 | 0.99 | 46 | 0.94 | 1.00 | 0.97 | 46 |
| Exhibition | 1.00 | 0.33 | 0.50 | 39 | 1.00 | 1.00 | 1.00 | 39 | 1.00 | 1.00 | 1.00 | 39 |
| Pink noise | 0.55 | 0.94 | 0.70 | 17 | 0.00 | 0.00 | 0.00 | 17 | 0.00 | 0.00 | 0.00 | 17 |
| Restaurant | 0.59 | 1.00 | 0.74 | 32 | 0.00 | 0.00 | 0.00 | 32 | 1.00 | 1.00 | 1.00 | 32 |
| Street noise | 0.40 | 0.33 | 0.36 | 48 | 0.00 | 0.00 | 0.00 | 48 | 0.98 | 0.94 | 0.96 | 48 |
| Subway | 0.54 | 1.00 | 0.70 | 32 | 1.00 | 1.00 | 1.00 | 32 | 1.00 | 1.00 | 1.00 | 32 |
| Train | 0.92 | 0.79 | 0.85 | 42 | 0.56 | 0.83 | 0.67 | 42 | 1.00 | 1.00 | 1.00 | 42 |

In Figure 4.37, confusion matrices are presented that were prepared for all tested models.




Naive Bayes classifier | Linear SVM | SVM polynomial

Gaussian Process | Decision Tree | Random Forest

MLP Classifier | AdaBoost Classifier | QDA

*Figure 4.37 Confusion matrices for all tested models*

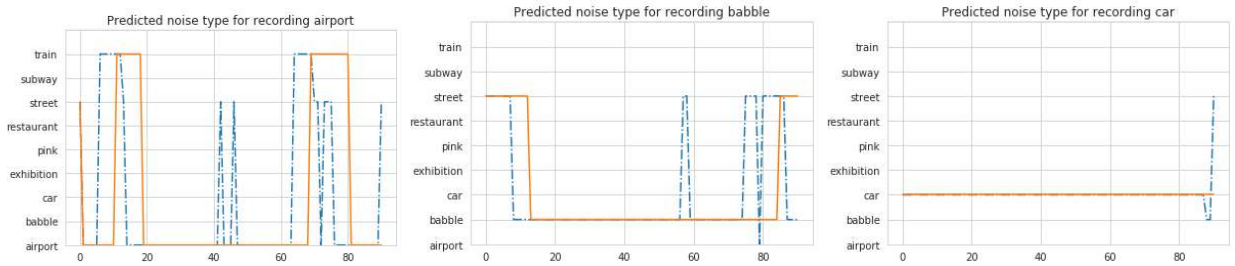
### 4.3.5 Discussion

The created model was tested on recordings that were used for training (but different parts of these recordings) and on the additional recordings from the multimodal corpus of English speech recordings called MODALITY (Czyzewski *et al.*, 2017). As mentioned before, in the context of noise profiling, the model's usefulness is measured by evaluating its stability, not

correctness. This is because the recording conditions might be very different - such as the recording method and equipment, sources of noise, and its characteristics. Therefore, for instance, the airport recording might be identified as street noise. What matters here is that this recording is always (or almost always) identified as street noise. That is why the correctness of classification is less of importance in general. The value of this model is in recognizing the abstract type of distortion using its frequency parameters − and this is the basis of improving speech intelligibility in the presence of noise. The process of speech quality/intelligibility enhancement requires particular conditioning − and the values of the parameters used should correspond to the type of noise. These values strongly impact the efficiency of speech intelligibility improvement. So, it is crucial to effectively classify the particular types of distortion to an assigned number of classes, enabling to modify speech in the best way in given noise conditions.

The recognition process was carried out in two modes: momentary and averaging. In both modes, the window/frame analyzed was 1 or 2 seconds, and the window was moved by 0.1 seconds with every step. In the momentary mode, classification was performed for every frame. In the averaging mode, the classification was made with delay - it means that the momentary classification should change across five subsequent frames to calculate the average classification. Thanks to this procedure, the recognition model avoids a temporary disturbance, usually caused by non-stationary noise.

Figures 4.38 and 4.39 present the outcomes of classification. The solid line represents the classification in the averaging mode, while the dashed line represents the momentary classification. The classification results for 1- and 2-second frames are different - first of all, it is because the learning process was performed using a 2-second frame; what is more, a longer window allows for better evaluation of the statistical features of the frequency characteristics. When using 2-second windows, the classification results are very good. For a 1-second window, the statistical features might not be visible so clearly, but the averaging mode provides satisfying results.



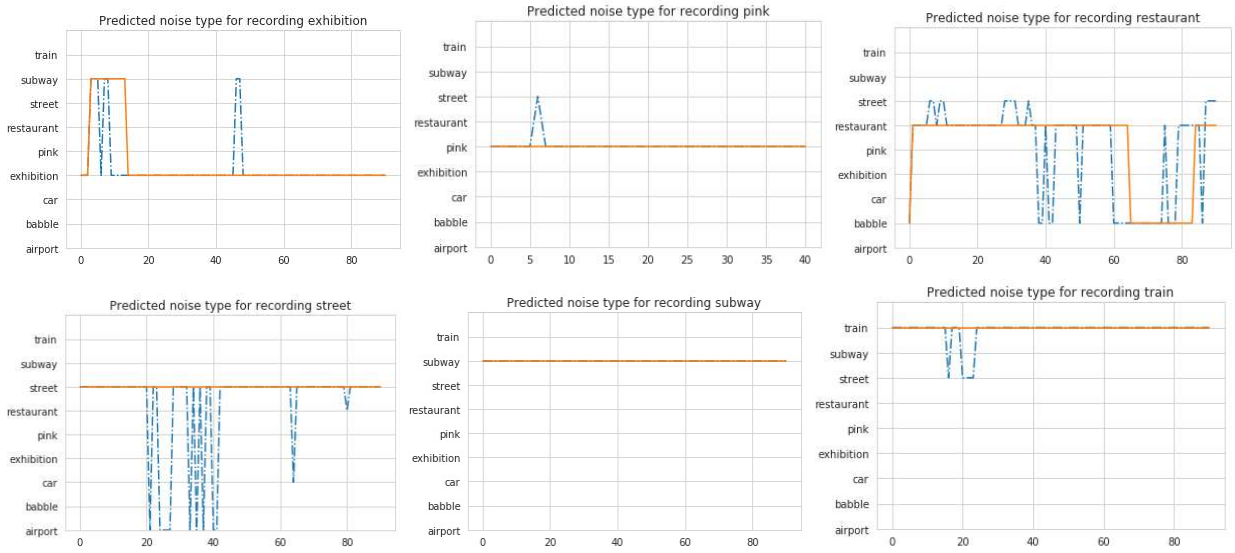



*Figure 4.38 Classification results on the real-life recordings using a 1-second-length frame*

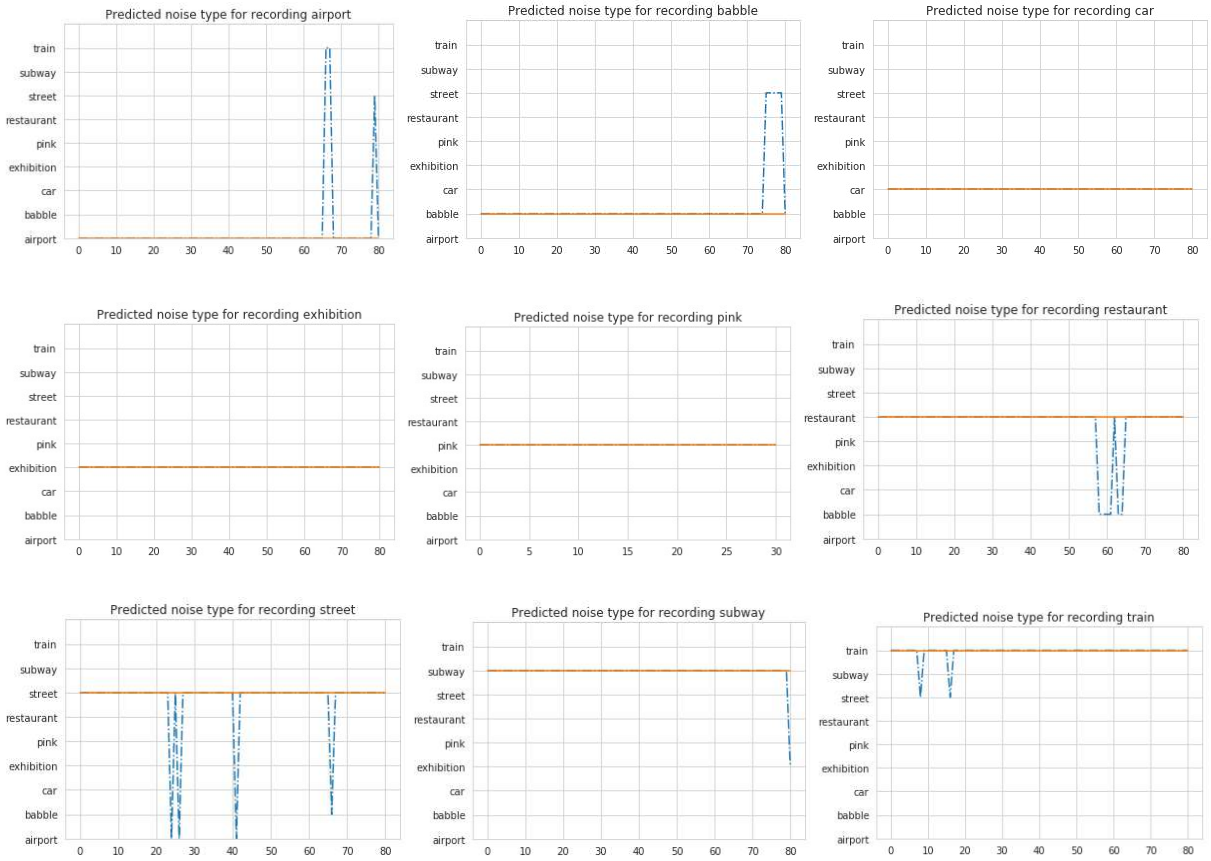



*Figure 4.39 Classification results on the real-life recordings using a 2-second-length frame*

The recognition process was also performed on a completely different set of noise recordings contained in the MODALITY (Czyzewski *et al.*, 2017) multimodal corpus of English speech recordings. Recordings used in this test were very long (between 11 minutes 45 seconds and 14 minutes 54 seconds). The test was performed only for a 2-seconds frame, and the window was moved by 2 seconds (due to the overall recording length) with every step. The averaging was also used to remove random fluctuations in the recognition results. Figures 4.40 to 4.42 present the results of recognition, with dashed lines representing the single window classification and the solid line representing the averaged result.

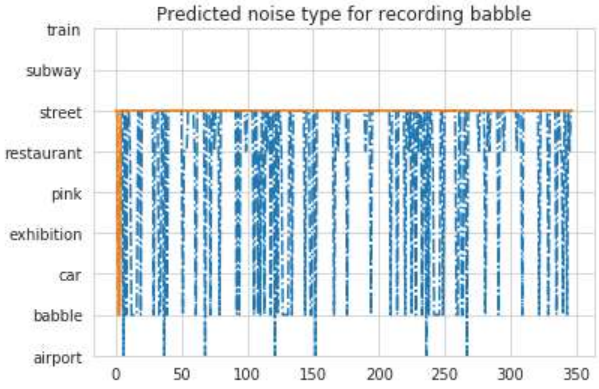



*Figure 4.40 An example in which the classification model has selected both "street" and "babble speech," but after averaging, the resulting classification was "street"*

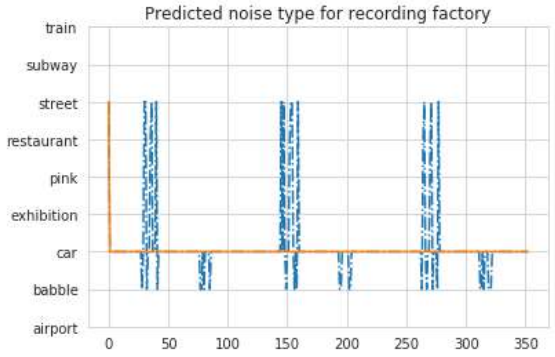



*Figure 4.41 An example in which the "factor" recording was classified as "car noise" (there was no such class as "factory" in the training set)*

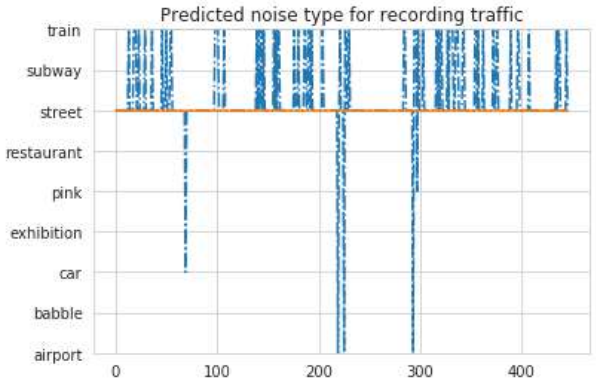



*Figure 4.42 An example in which the recording "traffic" was classified as "street," which is the correct classification*

As pointed out, it must be underlined that the classification quality is impacted by the stability of the classification, not correctness. That is why the results are generally satisfying even if the noise recordings are not correctly classified. As previously mentioned, classification will strongly be impacted by the recording place, recording equipment, sampling frequency, etc.

This way, thesis no. 2 has been proved.

## 4.4 Adaptive approach to speech modification

This chapter will cover the most crucial part of this dissertation – experiments performed to confirm the ability to build the system modifying the speech signal adaptively – to obtain the best possible improvement in speech quality. Figure 4.43 presents the current Chapter topic highlighted.

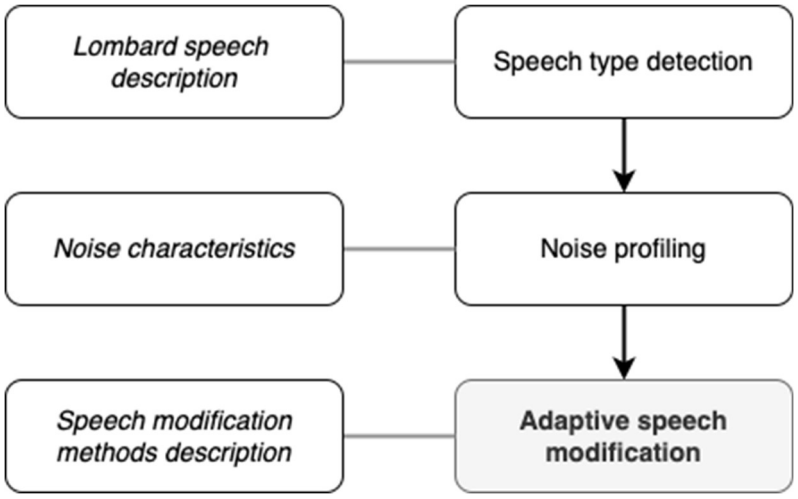



*Figure 4.43 Structure of the dissertation with current Chapter topics highlighted*

It was proven in Chapter 4.1.3 that modifying the fundamental frequency and formants can improve overall speech quality (and intelligibility). Moreover, there is a strong relationship between the nature of these changes and the speaker's attributes. According to the experiments performed, gender has the biggest impact on the effectiveness of such changes.

Considering the above, the experiments were performed to verify whether it is possible to find the most effective method of speech modification in terms of ensuring the highest speech quality results. Graphs in Figure 4.44 and Figure 4.45 show the average values of P.563 for changes performed on speech mixed with noise at the SNR=10 dB.

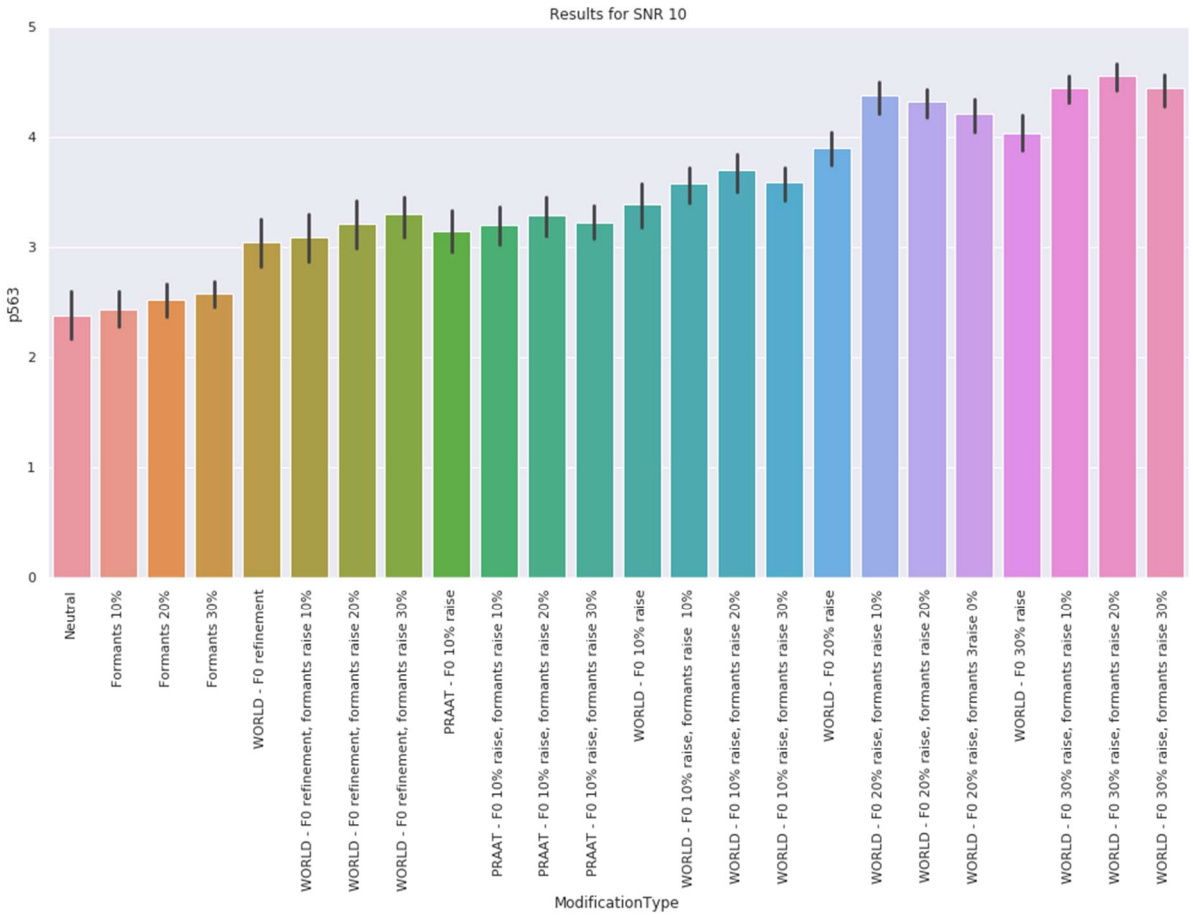



*Figure 4.44 Male recordings, various types of noise at SNR=10 dB*

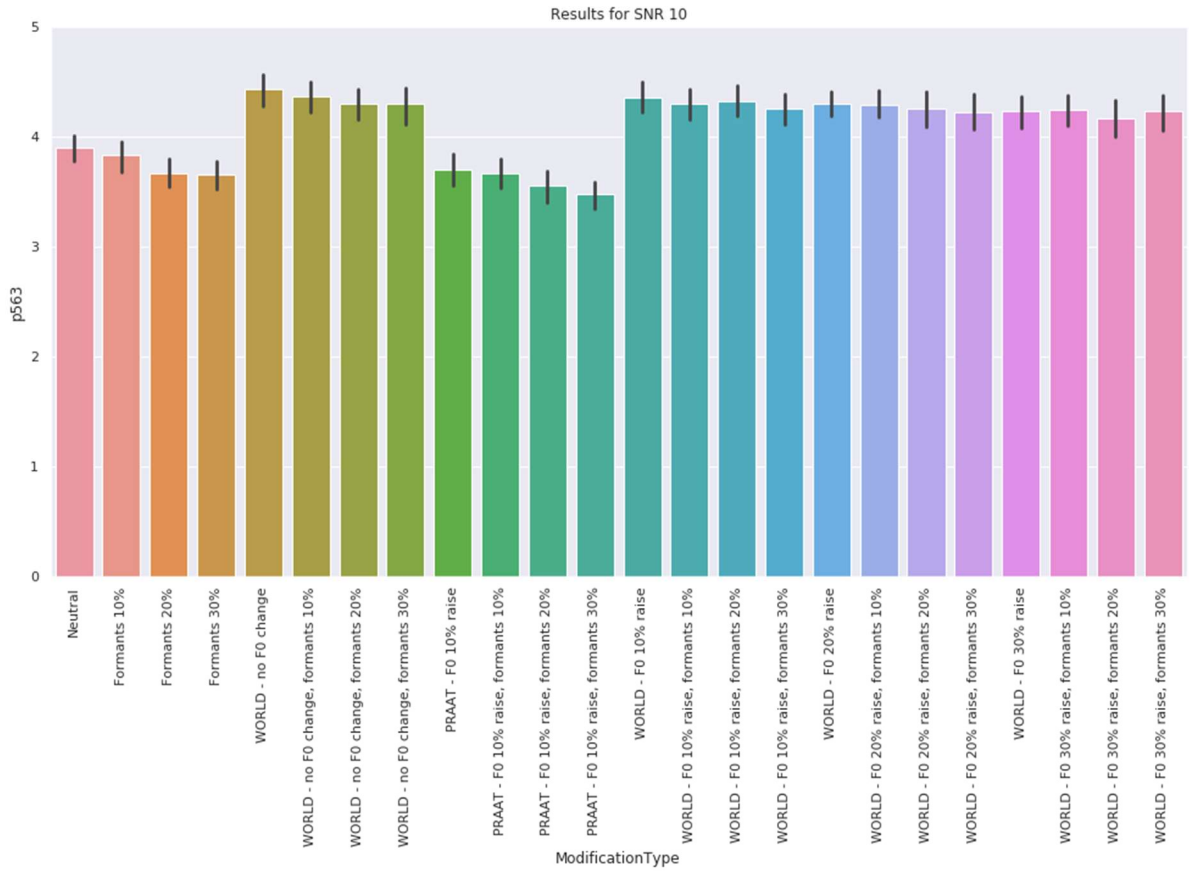



*Figure 4.45 Female recordings, various types of noise at SNR=10 dB*

It is clear that female speech needs to be modified in a completely different way - the best results were obtained for no change in the fundamental frequency but with a refinement of F0 only. For female voices raising F0 did not improve the speech quality almost at all.

Another situation was for male voices. Raising F0 (with its refinement) greatly improved the P.563 calculated value. When statistical values were calculated, it was, however, unclear what type of modification would be the best one. It appears that sometimes raising formant values helps a lot, but sometimes makes the overall quality worse. What is more, the type of noise and SNR value also have an impact on the calculated best values.

When the recordings were analyzed one by one, it appeared that every recording behaved a little differently, while the overall tendency to raise F0 (at male recordings) and only to refine it (for female voices) was maintained. The difficulty of using this logic is that there is a need to detect gender and use only two methods of speech modification. Detecting gender is possible; however, the gender detection methods focus on the detection process and not on the sound attributes that might impact the quality improvement metrics. In other words, these gender

detection methods were designed for a completely different purpose. In addition, the gender detection process is something that adds complexity to the system.

Therefore, it was assumed that gender would not be detected in the process of adaptive speech signal modification. To be able to define the best modification method for a given speech recording, the other approach was used.

### 4.4.1 Defining the best modifications for recordings

The first step in the process was to define the best modification available for the given recording. For that purpose, every recording was transformed and verified in the following way:

1. The neutral recording was modified using all the proposed methods (combination of methods):
    1. modifying F1-F3 formants by increasing their values by 10, 20, and 30%
    2. refining F0 by using WORLD vocoder (on clean speech); refining was also used along with no changes in formants, and with 10, 20, and 30% increased formants' values
    3. increasing F0 value using PRAAT by 10%; additionally, formants were left unchanged or raised by 10, 20, and 30%
    4. increasing F0 value by 10, 20, and 30% using WORLD vocoder (with F0 refinement) with no formants' change or increasing it by 10, 20, and 30%
2. For every previously generated recording (including the neutral speech) all noise types presented previously were added to the signal, with SNR levels of -5, 0, 5, 10, 15, and 20 dB. Lower values than -5 dB were not used since the speech is non-understandable at these levels mostly, and objective metrics (like P.563) do not return values higher than 1 for such signals (except when pink noise is used.)
3. Then the P.563 value was calculated for each of the recordings. Since there are 24 possible modifications, 6 SNRs, and nine noise types, the resulting dataset has 1296 recordings for one input recording of a neutral speech.
4. For each combination of input recording, SNR, and noise type, the best modification was defined (best modification means the modification that resulted in the highest P.563).

Sample results for a single file zd2_1 – a short sentence said by a male speaker are contained in Table 4.21.

*Table 4.21 Sample results for a short sentence uttered by a male speaker*

| | name | noise | snr | gender | ModificationType | p563 |
|---|---|---|---|---|---|---|
| **20413** | zd2_1 | car | 0 | m | WORLD - F0 20% raise, formants raise 10% | 3.458468 |
| **19657** | zd2_1 | car | 0 | m | WORLD - F0 20% raise | 3.238652 |
| **20575** | zd2_1 | car | 0 | m | WORLD - F0 30% raise, formants raise 10% | 3.189091 |
| **20251** | zd2_1 | car | 0 | m | WORLD - F0 10% raise, formants raise 10% | 3.016421 |
| **19711** | zd2_1 | car | 0 | m | WORLD - F0 30% raise | 2.693498 |
| **19603** | zd2_1 | car | 0 | m | WORLD - F0 10% raise | 2.420041 |
| **19441** | zd2_1 | car | 0 | m | Neutral | 1.791524 |
| **19927** | zd2_1 | car | 0 | m | WORLD - F0 refinement, formants raise 10% | 1.376786 |
| **19495** | zd2_1 | car | 0 | m | WORLD - F0 refinement | 1.321462 |
| **19765** | zd2_1 | car | 0 | m | Formants 10% | 1.000000 |

The best modification for this recording mixed with "car" noise on SNR = 0 dB is apparently F0 refinement with 20% raised F0 and formants increased by 10%. This modification was then marked as the best modification for this recording in the given conditions (SNR=0 dB, noise=car).

Every available recording from the corpus (Czyzewski *et al.*, 2017) was processed and verified in the same way to obtain a total of 77760 samples in the input dataset.

### 4.4.2 Limiting the number of best modifications

The target model is meant to perform classification. Because there are 23 potential modifications, it would result in poor overall accuracy (too many classes). What is more, some modifications resulted in unnatural speech - for instance, changing formants more than 10% resulted in a "childish" speech effect. Raising F0 with the PRAAT formula, on the other hand, did not usually improve the speech quality in noise, so this method was removed from the potential targets.

As a result, the following nine classes were defined:

1. Formants' values raised by 10% (marked as change "0X")
2. F0 refinement using WORLD with no formant change ("A0")
3. F0 refinement using WORLD with formants increased by 10% ("AX")
4. F0 increased by 10% with refinement using WORLD with no formant change ("C0")
5. F0 increased by 10% with refinement using WORLD with formants increased by 10% ("CX")

6. F0 increased by 20% with refinement using WORLD with no formant change (“D0”)
7. F0 increased by 20% with refinement using WORLD with formants increased by 10% (“DX”)
8. F0 increased by 30% with refinement using WORLD with no formant change (“E0”)
9. F0 increased by 30% with refinement using WORLD with formants increased by 10% (“EX”)

### 4.4.3 Gathering data for the ML model

The second step is to gather the data that would feed the model. The first attempt was to create the model based on the calculated values of fundamental frequency (F0) and the first three formants. But it led to a weak model with low accuracy (around 40%). The features used in the learning process were as follows:

1. F0
2. F1-F3 values
3. Noise type
4. SNR value

The label was defined as the best modification of the signal. However, the average values worked poorly since they did not differentiate between speech recordings. That’s why the approach was changed, and the following process was implemented to calculate the best modifications data for the ML model:

1. The model features were calculated for 1-second windows.
2. Fx averages were calculated only for the parts that have F0 values different from 0 (so only sound speech fragments were used).
3. The windows for which features were calculated were moved by 0.1 second forward (the hop length was 0.1 second).
4. The label (desired value) for the given frame was taken from the overall best modification for the given recording, noise, and SNR. It is because P.563 cannot be effectively calculated for such short recordings (duration less than 1 second; it was previously mentioned that P.563 metrics could be effectively measured for recordings longer than 3 seconds).

Since the results were still weak (around 42% of accuracy using an unlimited decision tree), another approach was undertaken:

1. Window definition remained the same (1-second window, 0.1-second hop length).
2. For every window, F0 was calculated along with the first 12 MFCCs.

This results in a set of multiple windows (fragments) for a single recording. The overall number of records in the dataset used to create the model was 52596.

### 4.4.4 Explanation of the models

**Decision tree model**

The idea of the model is to calculate the best modification for a given speech fragment. It is assumed that the best modification depends on the sound features (F0 and MFCCs), noise type (obtained by the automatic noise profiling mentioned earlier), and SNR (calculated directly).

As a first attempt toward building the final model, the decision tree was used. Below, in Table 4.22, the results of decision tree implementation are contained.

*Table 4.22 Results of decision tree model implementation*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0X | 0.70 | 0.75 | 0.72 | 168 |
| A0 | 0.79 | 0.80 | 0.80 | 1562 |
| AX | 0.73 | 0.70 | 0.72 | 718 |
| C0 | 0.72 | 0.72 | 0.72 | 976 |
| CX | 0.75 | 0.73 | 0.74 | 776 |
| D0 | 0.70 | 0.70 | 0.70 | 1311 |
| DX | 0.76 | 0.77 | 0.77 | 3226 |
| E0 | 0.76 | 0.75 | 0.75 | 1001 |
| EX | 0.86 | 0.87 | 0.87 | 6041 |
| accuracy | | | 0.79 | 15779 |
| macro avg | 0.75 | 0.75 | 0.75 | 15779 |
| weighted avg | 0.79 | 0.79 | 0.79 | 15779 |

ROC: 0.8621210843309045

The overall accuracy of this model is 79%, which is sufficient for practical use. The confusion matrix gives some more insight into the model accuracy (see Figure 4.46).

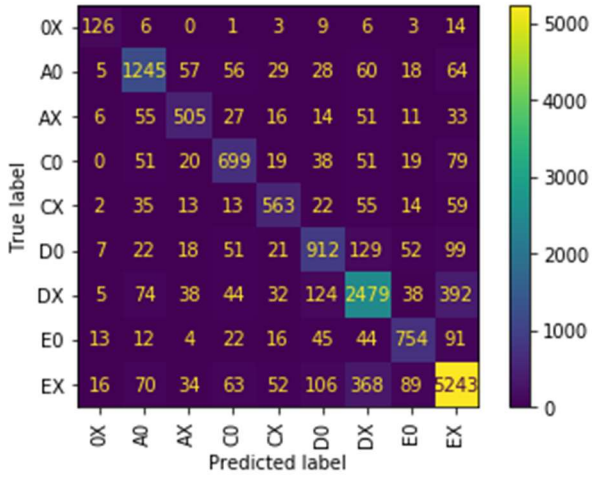



*Figure 4.46 Confusion matrix for decision tree model used in adaptive speech improvement*

The importance of features in this model is presented below in Figure 4.47.

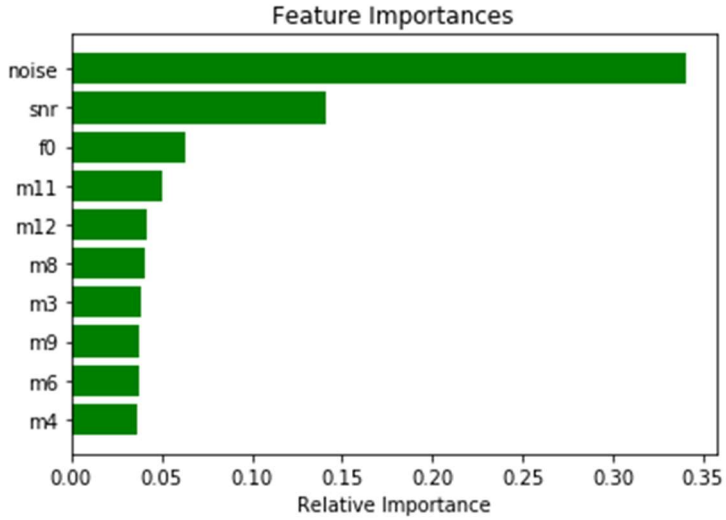



*Figure 4.47 Feature importance ("m[x]" features are the given MFCCs)*

The only drawback of this approach is that the created tree depth is 28 - which is a lot and indicates that the model is probably highly overfitted. The only way to limit the overfitting would be to determine the maximum depth. But the resulting tree has very low accuracy.

**Multi-layer perceptron (MLP) classifier**

The second attempt was to build an MLP classifier. The structure of a neural network is as follows:

1. There are three hidden layers
2. Every hidden layer has 100 neurons

3. The optimizer chosen is Adam
4. The number of epochs is 600

The test set accuracy is 72.5%, and the loss is 0.48. Below, the results of the training phase are presented:

- Training set score: 0.82
- Testing set score: 0.73
- Training set loss: 0.48

Detailed metric values for MLP are shown in Table 4.23.

*Table 4.23 Results of the MLP classifier implementation*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0X | 0.67 | 0.69 | 0.68 | 168 |
| A0 | 0.74 | 0.71 | 0.73 | 1562 |
| AX | 0.67 | 0.69 | 0.68 | 718 |
| C0 | 0.70 | 0.52 | 0.60 | 976 |
| CX | 0.58 | 0.67 | 0.62 | 776 |
| D0 | 0.65 | 0.46 | 0.54 | 1311 |
| DX | 0.66 | 0.71 | 0.68 | 3226 |
| E0 | 0.74 | 0.56 | 0.64 | 1001 |
| EX | 0.79 | 0.87 | 0.83 | 6041 |
| accuracy | | | 0.73 | 15779 |
| macro avg | 0.69 | 0.65 | 0.67 | 15779 |
| weighted avg | 0.72 | 0.73 | 0.72 | 15779 |

ROC: 0.95823017129792

The confusion matrix is presented below (see Figure 4.48).

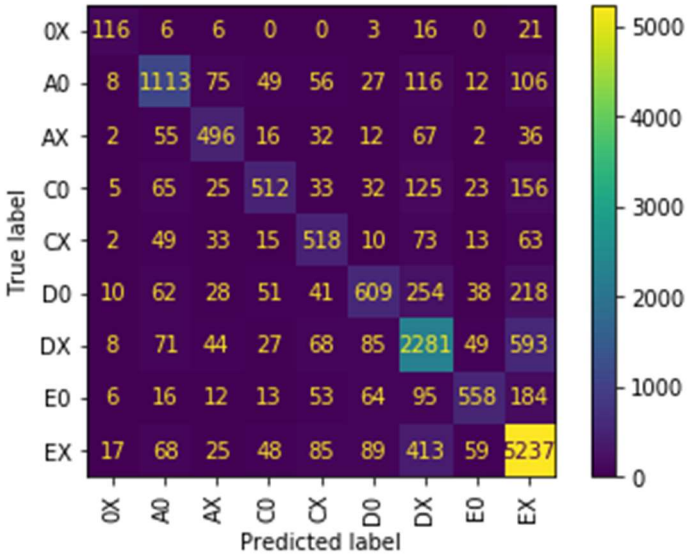



*Figure 4.48 Confusion matrix for the MLP model*

In Figure 4.49, the training loss curve is shown.

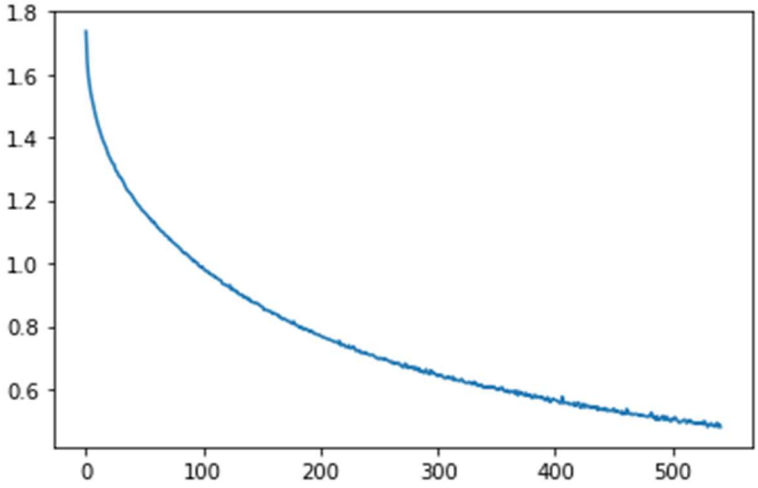



*Figure 4.49 Training loss curve for the MLP model*

The idea of inference is the same as the learning process and preparation of the dataset: the input signal is divided into small 1-second fragments, and the best modification is selected based on the above model. The overall results of the adaptive improvement process are presented in Chapter 5.

# 5 Evaluation of results

## 5.1 Objective

The adaptive modifications performed on the input speech signal result in improved speech objective quality measures. But the typical signal changes (like increasing F0 by 10 or 20%) provide the desired improvements. The possible way of modifying the input speech signal would be to increase F0 by 10% and formants by 10% as a static, typical modification. As previously proved, different recordings have different optimal signal changes (where optimal means the modification that leads to the best P.563 value).

That's why the adaptive signal modification process is proposed. A comparison of the adaptive modification will be made versus the typical signal changes. The adaptive changes should give at least similar or better results than the typical ones. Adaptive modification is calculated using the MLP classifier described in Chapter 4.

The inference is made using the following process:

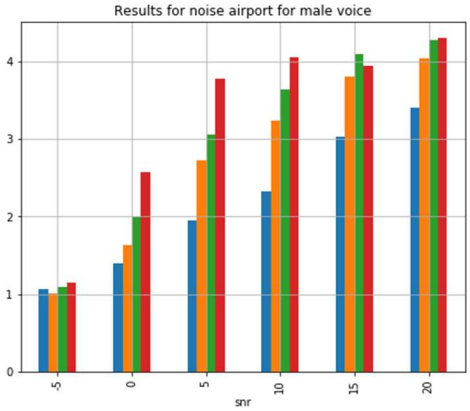
1. F0 values are calculated for the given recording (the resulting vector has F0 values every 80 samples; the sampling frequency is 16kHz).
2. MFCCs are calculated for the given recording (the resulting matrix has the same time resolution as the F0 vector).
3. Then the 1-second fragment is taken into account, and averages for F0 and the first 12 MFCCs in this 1-second window are calculated. As a result, for this 1-second window, 13 features are derived, i.e., average F0 value and average values of the first 12 MFCC coefficients.
4. To the previously calculated 13 features, noise type and SNR are added, thus forming the 15-element input vector. The classification is then performed on this vector.
5. The following window is calculated by moving 0.1 seconds forward. For every window, optimal modification is calculated and selected from the nine available classes mentioned earlier.
6. The final modification is selected by choosing the transformation that occurs most often for all windows. This works as selecting the median value for statistical distribution. It is a way of averaging the result since the classification does not work in real-time.

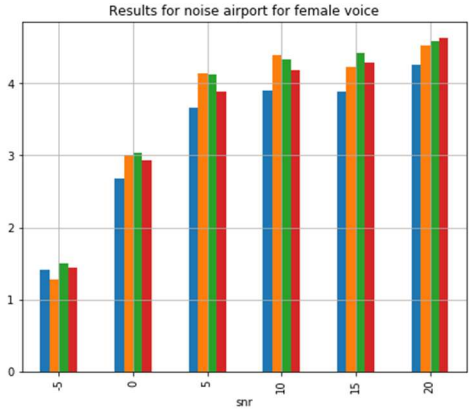
The results differ depending on the speaker's gender. For the male speakers, the adaptive modifications give much better results. For female speakers, the differences are very small; the

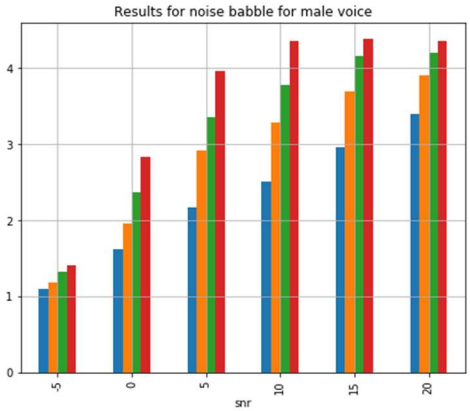
adaptive changes are sometimes a little worse than the selected typical changes. This is, however, a trade-off because the adaptive way does not care about the speaker's gender. It just takes the F0 and MFCCs into account, which greatly simplifies the pipeline, and the results are overall better.

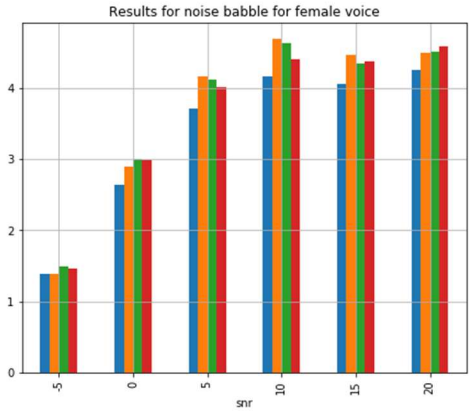
The results are calculated for only three types of modifications since not every modification provided promising results. So, the below charts (see Figure 5.1) compare the following types of speech signals:

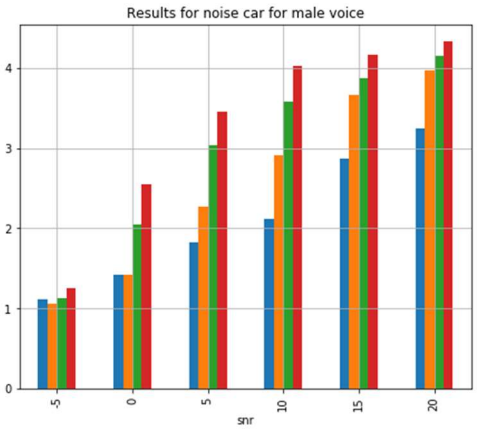
- neutral - speech without modification mixed with a given noise signal,
- standard_ax - speech with F0 refinement using WORLD vocoder, with formants increased by 10%,
- standard_cx - speech with F0 raised by 10% and refined, and with formants increased by 10%,
- adaptive - the result of the adaptive modifications.

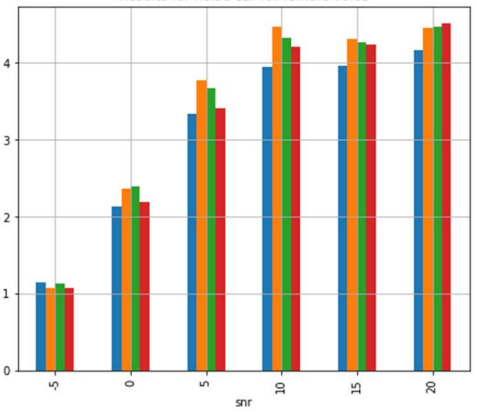


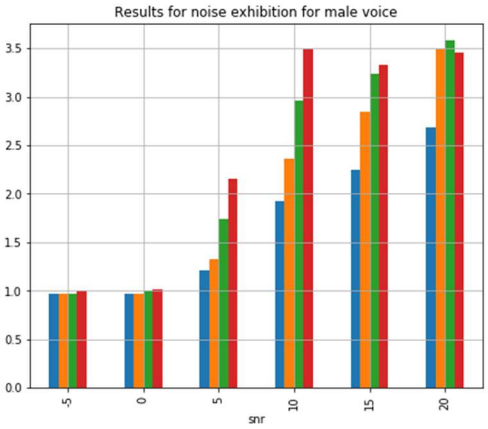



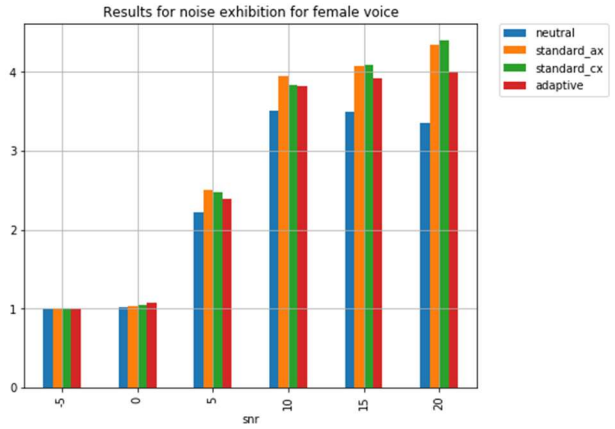



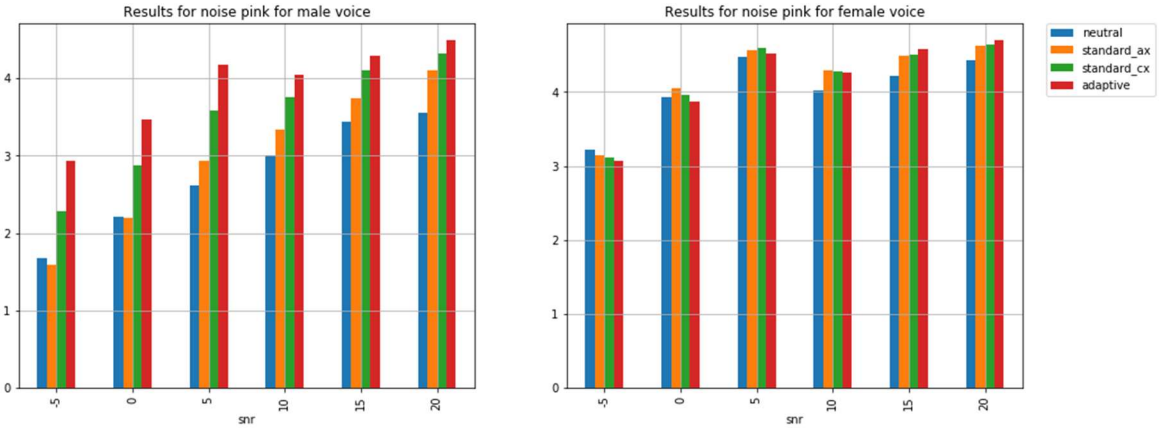



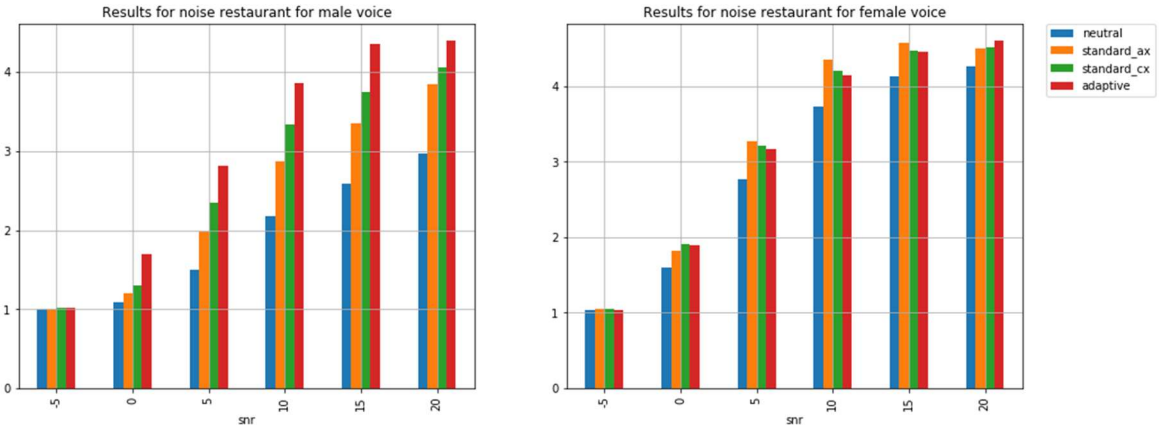



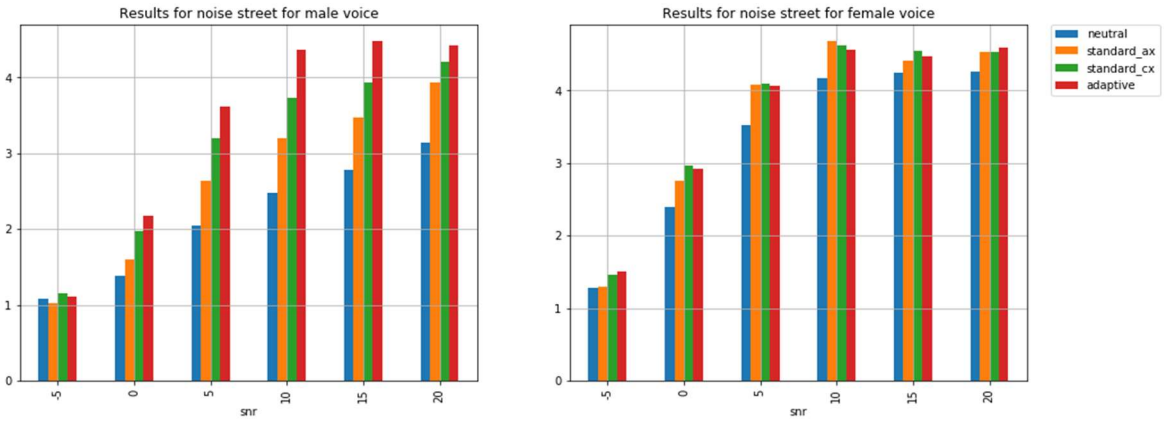



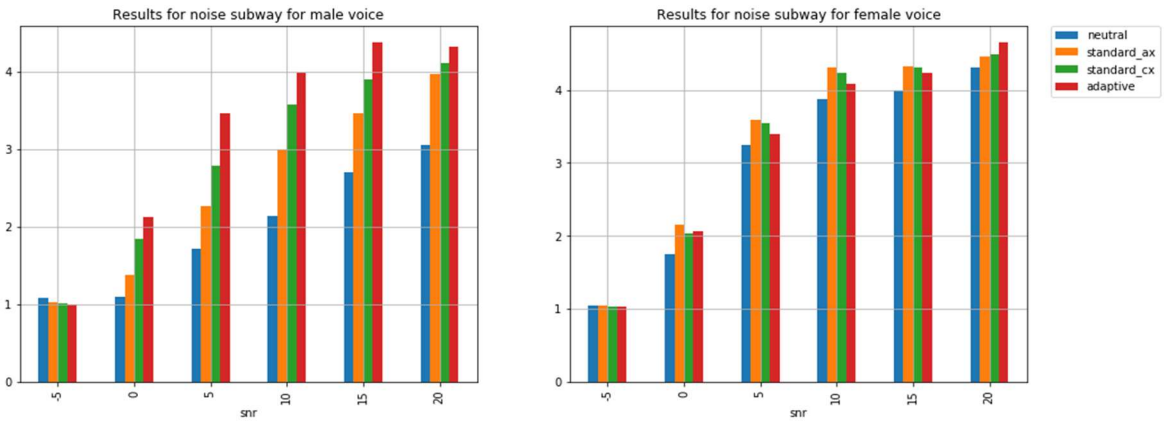





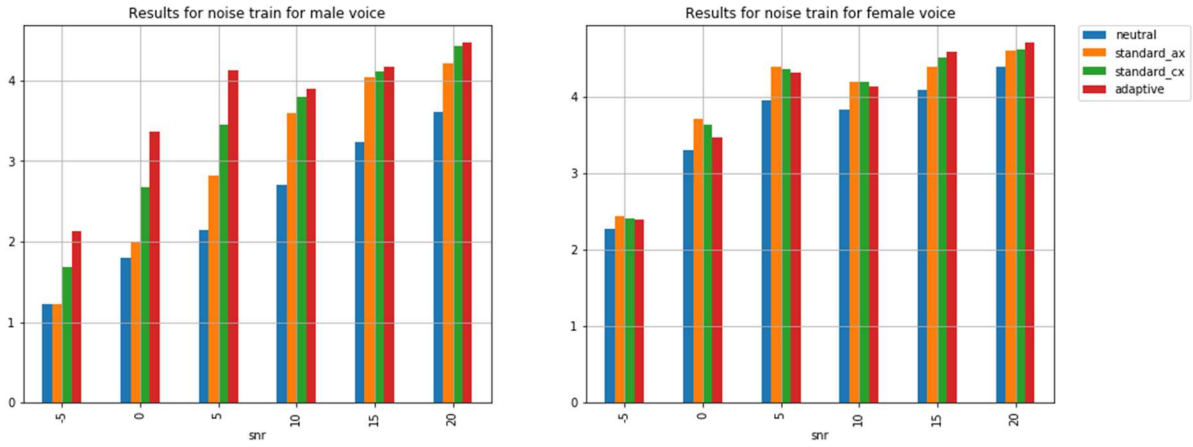



*Figure 5.1 Results of the experiments for different genders and noise types*

It is clear that it is much more challenging to improve the quality of the female voice, both using the predefined methods and adaptive process. A detailed comparison of the results of the improvement methods is presented in Appendix B.

## 5.2 ANOVA test

The ANOVA test is used to confirm two hypotheses:

- The adaptive speech modification results depend on noise type, which means that the noise should be profiled before the modification occurs.
- The results of the different types of modifications have different average values.

The first ANOVA test was performed for all adaptively modified recordings, separately for both genders and all SNR levels. First, the MOS values of P.563 were compared for each combination. For this purpose, the distribution plots of P.563 values presented in Figure 5.2 were prepared.

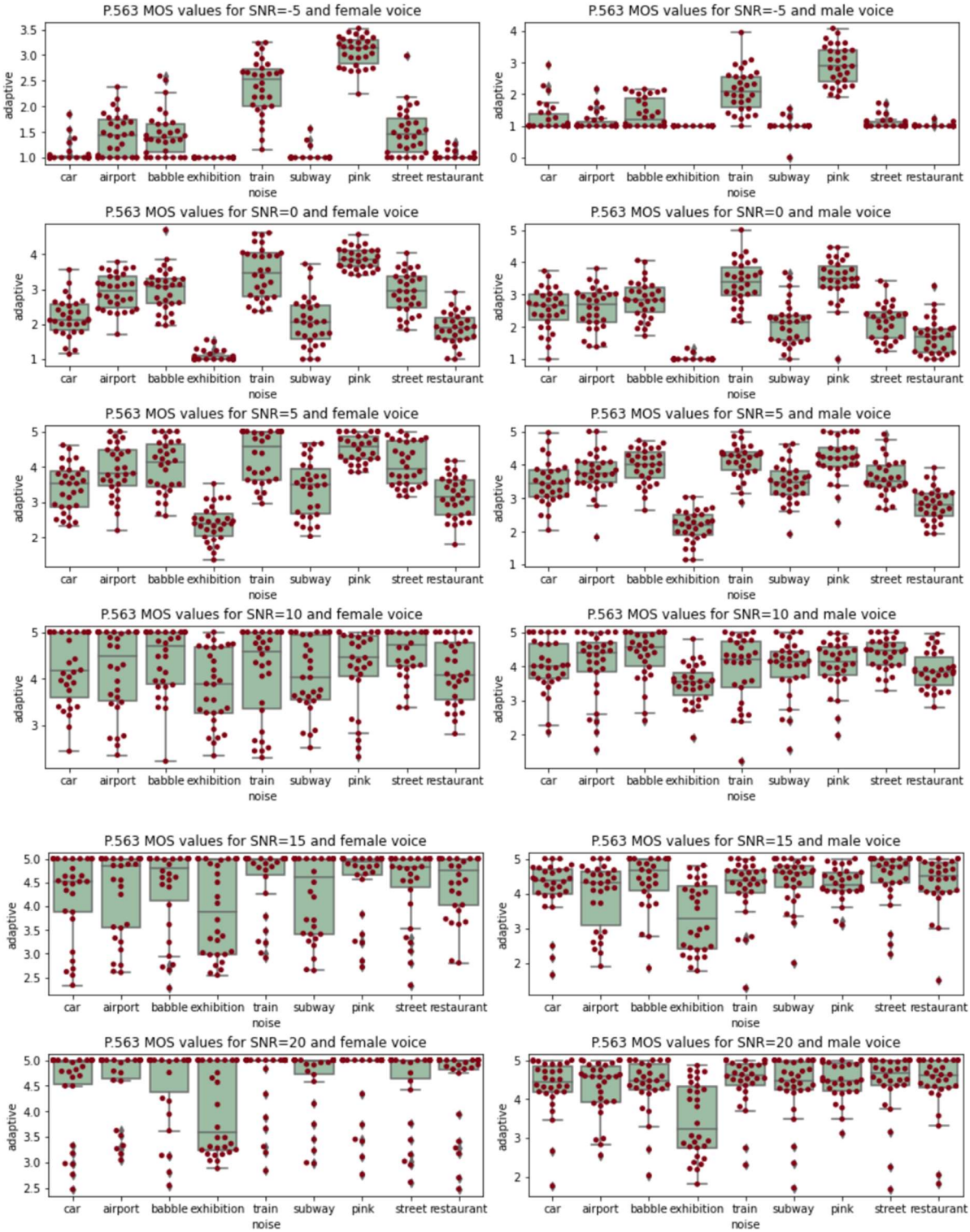



*Figure 5.2 MOS P.563 distribution for all SNR levels and genders by noise type*

It can be seen from Figure 5.2 that for lower SNR values, the P.563 levels depend strongly on the noise type. In contrast, for higher SNR values, the distributions are different, but it can be

hypothesized that the averages for most noise types are similar. This is also confirmed by the results of the ANOVA test, presented in Table 5.1.

*Table 5.1 ANOVA test results calculated for MOS P.563 values in the adaptive model. The statistically significant values are highlighted in red ($\rho<0.05$)*

| SNR | GENDER | $\rho$ | F |
|---|---|---|---|
| -5 | Female | 4.85E-89 | 137.5929 |
| -5 | Male | 7.35E-67 | 82.1473 |
| 0 | Female | 1.91E-63 | 75.3729 |
| 0 | Male | 2.74E-49 | 51.2502 |
| 5 | Female | 1.51E-33 | 30.5737 |
| 5 | Male | 1.77E-41 | 40.2845 |
| 10 | Female | 0.04459571 | 2.0188 |
| 10 | Male | 0.00040396 | 3.7059 |
| 15 | Female | 0.05535124 | 1.9339 |
| 15 | Male | 7.02E-07 | 5.8475 |
| 20 | Female | 0.01167245 | 2.5235 |
| 20 | Male | 2.07E-06 | 5.4853 |

The second ANOVA test was performed for the second hypothesis. The relevant F values were calculated for the whole set of results using the type of modification as a variable. Then the F-values were calculated for all types of noises separately. In Figure 5.3, all the distributions were presented. The overall F-value for this test equals **15.63,** and it is statistically significant. When

considering all noise types separately the statistical significance is presented in Table 5.2. The statistically significant values are highlighted in red ($\rho$<0.05).

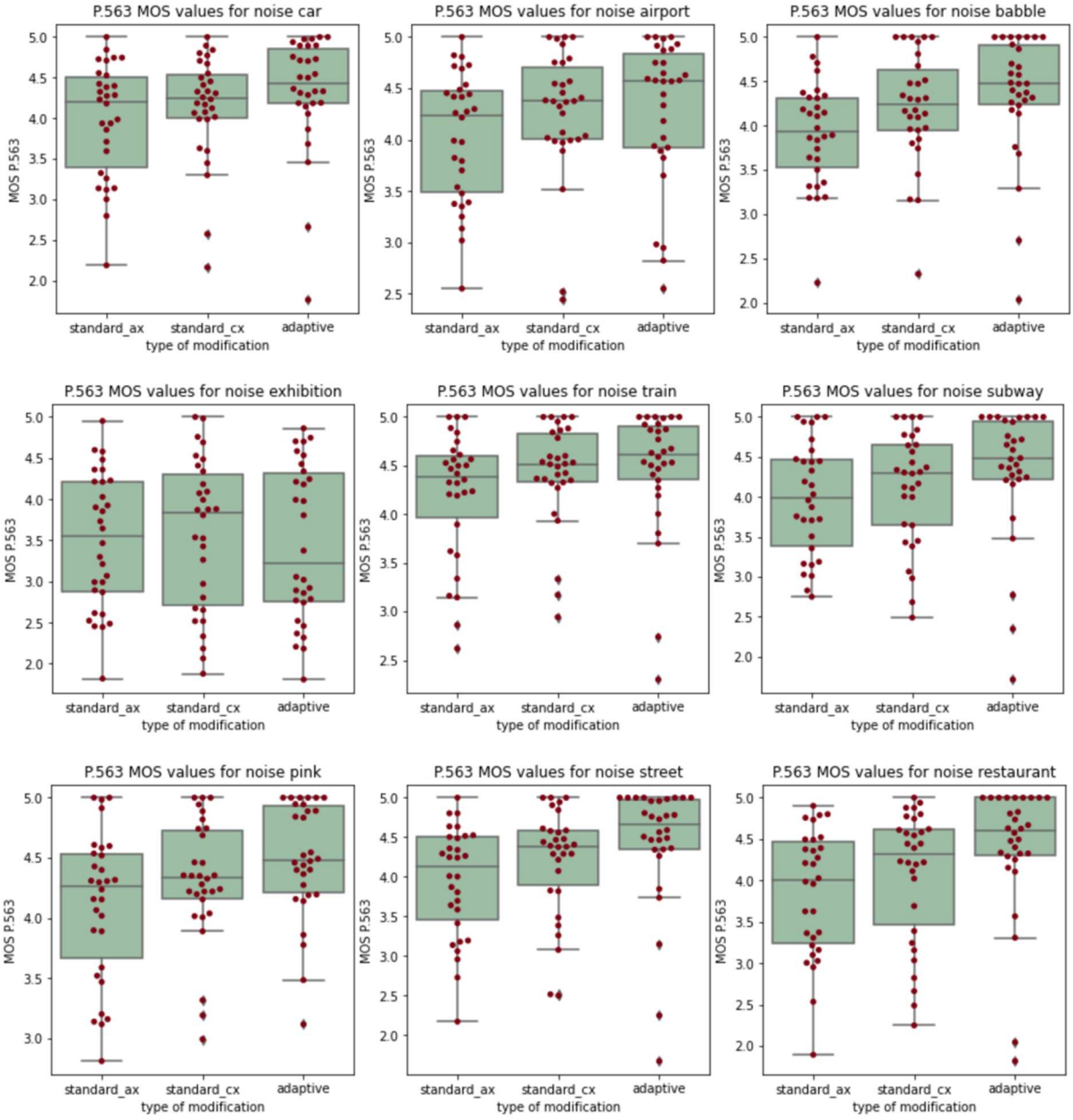



*Figure 5.3 MOS P.563 distribution for the different types of noises and speech modifications*

*Table 5.2 ANOVA test results for different types of noise*

| NOISE | $\rho$ | F |
|---|---|---|
| Car | 0.1507 | 1.9343 |
| Airport | 0.2311 | 1.4898 |
| Babble | 0.0268 | 3.7733 |
| Exhibition | 0.8632 | 0.1474 |
| Train | 0.2148 | 1.5656 |
| Subway | 0.1903 | 1.6913 |
| Pink | 0.0272 | 3.7582 |
| Street | 0.0394 | 3.3565 |
| Restaurant | 0.0310 | 3.6158 |

## 5.3 Effectiveness of the method in terms of computation time

One of the most important factors allowing for judging the effects of the adaptive speech modifications is the computation efficiency. The system consists of multiple components that might result in signal delay. However, the selected methods were chosen with performance in mind. What is more, some of these components do not have to work in continuous time – for instance, noise detection does not have to be performed continuously.

It is absolutely critical that the detection and modification algorithms are as performant as possible. To simulate the real environment conditions, time measurements were performed on a typical business-type laptop with Ubuntu 20.04 LTS operation system installed. The device parameters are as follows:

- CPU: Intel Core i5-6300U, 2.40 GHz x 4 cores
- Built-in graphics adapter (no possibility to use GPU acceleration for neural networks inference)
- RAM: 8GB

Recalling measurement descriptions and the results obtained are described further on.

**Lombard speech detection**

Every recording available (120 recordings in Polish) was used in the inference process and treated as one inference entity – which means that the time was measured for the whole sound recording, even though the detection time was 1/3 of a second and step is also 1/3 of a second. This means that the windows overlap does not exist. In production systems, the detection might be less frequent since the speech signal character does not change so often. However, for the sake

of this experiment, the assumptions have been made that all parts of the signal must be covered by the system.

The average detection time for the recording is 1.2 seconds, while the average duration of the recording is 2.67 seconds. Most importantly, the detection time is – on average – 45% of the recording time. This means that the Lombard speech detection can be done in near real-time.

**Noise profiling**

Noise profiling measurements have been performed by just detecting the noise type using its spectral parameters and Naïve Bayes classifier.

By average, 1 second of the noise has been correctly classified in 5.83 ms.

**Calculating MFCCs and F0**

Since calculating the MFCCs and the fundamental frequency is crucial for the process of adaptive speech modification, it has also been measured how fast these parameters can be measured for the whole recording.

The average calculation time for the whole recording is 214 ms, which also proves that the time will not affect the ability to implement the near real-time system.

**Adaptation**

The adaptation time means the time needed to detect the best modification method. In other words, it is the time needed to detect the best modification using the precalculated MFCCs and F0 parameters. The process uses inference with the built neural network model. The average time of inference for a single speech frame (1 second) is 0.5 ms.

**Conclusion**

The most time-consuming process is the Lombard speech detection, which is obvious since the inference requires taking a virtual picture of the signal (the real picture is not generated and saved – it is not necessary to make an inference and would take a lot of time due to input/output process) and making a classification using the convolutional neural network.

All other processes are very fast; thus, the capability to work in near real-time has been confirmed. This way, thesis no. 3 has been proved.

## 5.4 Overall discussion

The experiments performed proved that the speech quality can be adaptively improved without gender detection. It is evident that in the real-time (or near real-time) systems, all the processes and transformations must work an as-fast-as and straightforward possible way to not add any delays to the system.

The proposed method is very fast and uses, in fact, simple models that do not require high computational capabilities to work. What is more, the models are further simplified by removing some – potentially necessary – components, like gender detection. As was shown, Lombard speech modifications are highly gender-specific, but gender detection requires an additional deep learning model that should have been included in the pipeline.

It is also worth mentioning that the system might be further simplified for some applications (for instance, by removing the continuous noise and speech profiling and doing both processes only periodically – to verify if the conditions have not been changed; the assumption that the conditions do not change too often is usually justified). The only component working continuously is then the speech improvement component which can work in near real-time, thus incurring only very small delays. This was already noted by the work of Morise et al. (Morise *et al.*, 2016).

# 6 Conclusions

## 6.1 Overall conclusions

Speech improvement in noisy conditions is a complex topic considering that this is only the clean signal processed in the channel. Therefore, it is challenging to change the signal in real-time without knowing its phonemic contents. What is more, the changes in the signal should be performed in real-time or with minimal delay.

The assisting noise may drastically decrease speech intelligibility. It can be best observed when the speaker is not aware of the noise – which means that usually, the speaker in such conditions does not use the natural mechanism of the Lombard effect.

The system capable of working in these conditions must be simple enough to provide the ability to work in a near real-time way and effective sufficiently to improve speech intelligibility as much as possible. There are many use cases for such systems, like broadcasting systems (since the broadcasting systems usually work in separated environments – where the speaker is not aware of the noise). Chapter 6.4 also proposes some other applications which could be developed using the proposed system's approach.

The proposed adaptive system uses a couple of important components:

- Speech type detection – to avoid changing the attributes of Lombard speech,
- Noise profiling – to be able to adapt the improvement method,
- Adaptive system of speech modification.

All of these components work using machine learning algorithms to provide state-of-the-art quality. The algorithms are different for all components, and the best choices were selected using the performance and simplicity factors:

- Speech type detection works as a relatively simple convolutional neural network,
- Noise profiling is based on spectral features of noise, and noise is classified using averaging on the Naïve Bayes model output,
- The process of adaptive selection of the best speech modification was implemented using a simple neural network classifier (Multi-layer Perceptron).

The results are highly promising – since the system provides great improvement for male speech and is performing similar to the best, experimentally selected method of changing female voice. It is important to underline that the adaptive system does not detect the speaker's gender, which positively impacts the overall system's simplicity and performance.

## 6.2 Proving theses

This dissertation was created with the following theses to prove:

**1. Employing a Convolutional Neural Network (CNN) enables the detection of the Lombard effect in uttered speech with sufficient accuracy for the purpose of speech profiling.**

Convolutional Neural Networks have usually been used for image classification challenges, but they are also employed for sound classification tasks, such as recognizing speech, music, environmental sounds, etc. In this dissertation, it has been proven that it is also efficient in Lombard speech detection. Of course, the accuracy is better when the gender is known in advance, but even when it is unknown, the accuracy reaches as high as 80-85%, which is a good result in terms of classification, especially as it is almost impossible to detect this effect by other methods.

CNN trained to detect the type of speech is relatively simple; thus, the inference is fast enough to be a part of the near real-time pipeline.

**2. Baseline machine learning methods can be used to effectively profile the ambient noise in a stable and near real-time manner.**

Stable noise type recognition is crucial for the speech improvement adaptative model. It was proven that baseline machine learning algorithms, based on the spectral characteristics of the noise signal (spectral bandwidth, spectral centroid, and spectral flatness used along with their statistical parameters), allow for stable and fast noise type detection.

The accuracy of the detection model, when used with an averaging algorithm, is very high, and the model itself is sufficiently accurate to provide predictable information for the adaptive algorithm.

**3. It is possible to build a system that adaptively modifies the speech signal to improve the speech intelligibility in noise optimally based on speech and noise profiling.**

Adaptive speech quality improvements, using the Lombard effect, have been tested by many researchers using the TTS systems (Bollepalli *et al.*, 2019; López *et al.*, 2017). It has been proved in this dissertation that it is possible to improve speech quality even when the phonemic content of the speech is unknown.

The adaptive speech quality improvement allows for better and faster results without the necessity to recognize gender and examine any complex parameters of the speech. What is more, the WORLD vocoder is used to both get the fundamental frequency and synthesize the speech with changed F0. This simplifies further the experiment pipeline.

It has been proven that the adaptive system might work efficiently, with low inference speed and a real-time vocoder (WORLD).

## 6.3 Achievements, contributions to the area of interest

There are many works in the scientific world that treat the Lombard effect as a method of improving the speech quality for text-to-speech systems. However, these systems work in other conditions than this dissertation considers. Text-to-speech systems have the textual input – thus, they are aware of the phonemic and – quite often – semantic and syntactic contents of the text.

Semantic and syntactic content is critical for text-to-speech systems since generating the natural speech requires implementing the prosody flow, which is only possible if the system is aware of the syntactic and semantic content.

From the Lombard effect generation perspective, it is, however, critical to know the phonemic contents – it simplifies the system and enables much more modification methods (like vowels time extension). These systems work quite well, proving that the Lombard effect is helpful for such AI-based systems.

This dissertation covers another area of speech quality improvement – when no phonemic, semantic or syntactic content is provided. Then the speech is just a sound signal, which quality and intelligibility have to be improved. Of course, it is much more difficult than in the text-to-speech systems since no phonemes can be recognized at speed allowing for near real-time applications.

This work proves the ML-based way to improve speech quality and intelligibility and proposes further system development that might be used in various applications. Most importantly, this dissertation proposes using adaptive speech quality improvement based on noise and speech features.

## 6.4 Future direction

In terms of future development, the system might be improved in a way that will extend its potential usage areas. Its main feature is that the clean speech signal is available and is separated from the noise. When speech is recorded with noise, the fundamental frequency, and formant manipulations become more complex and risk signal quality degradation.

This limits the current usage of the system to only some scenarios – like broadcasting systems. Although it is probably impossible to create a single system that will work in any conditions, the proposed approach might be helpful in designing other similar systems that might effectively work for different use cases.

There are particular areas for which the system might be developed and used:

**Using adaptive speech intelligibility improvements in hearing aids**

People using hearing aids usually encounter problems with speech intelligibility in noise. Even though the modern hearing aids have advanced noise cancellation implementations – the signal will always contain distortions. Therefore, it would be useful if there was an adaptive system that could detect the type of noise and – using the Lombard effect approach – improve the speech quality.

Such a system could work in the following way:

1. First, the Voice Activity Detection system should be used to detect the signal not containing speech.

2. This could then be used to detect the type of noise.

3. Adaptive system can be utilized to make a proper adaptation of sounding parts of the input signal, but without the usage of vocoders, to avoid signal quality degradation when no clean speech is available.

These types of systems could greatly improve the hearing aids users' experience. To support such a notion, a paper published recently (Saba and Hansen, 2022) can be recalled.

**Public speaking systems**

In public speaking events, the noise comes from the audience (usually). So, there is a strong feedback loop in the sound channel. The adaptive system could improve how the voice is transmitted through such channels and enhance the listening experience.

Such a system should work in the following way:

1. Noise type detection through the same microphone the speaker is using.
2. Noise cancellation in the channel.
3. Adaptive speech quality improvement, using the same approach as proposed in this work.

**Voice Conversion**

Recently, new trends appeared in voice conversion (Lee *et al*., 2021; Luong, Tran, 2021; Łatka *et al*., 2019; Merritt *et al*., 2022). Among them is the so-called many-to-many voice conversion (Lee *et al*., 2021; Luong, Tran, 2021; Merritt *et al*., 2022) based on deep models. Specifically, variational autoencoders (VAEs) are used to disentangle speaker identity and linguistic content from utterances. The method assumes that a deep model may convert from many source speakers to many target speakers with a single VAE (Luong, Tran, 2021). Thus, we can

envision that such an approach based on VAE may be employed for converting natural speech into Lombard speech.

Finally, there are other scenarios in which speech intelligibility is important, but – concurrently − noise is prevalent, e.g., emergency call centers (Gałka *et al*., 2015), workplace or call centers. If there is a need to communicate in such conditions via a public address system, one may envision that conversion into Lombard speech can benefit the recipients.

**Conclusion**

Speech quality and intelligibility improvements are a very important topic, and thus the effects of this work might be widely used in many typical use cases where speech is heard in the presence of noise. Using the Lombard effect brings gratifying results and is relatively simple to implement. What is most important is that it might work in near real-time, which in most such systems is a basic – but, at the same time – a fundamental requirement.

# 7 References

Abe, T., Kobayashi, T. and Imai, S. (1997), "The IF spectrogram: a new spectral representation," *Proc. ASVA 97*, pp. 423–430.

Adams, S.G. and Lang, A.E. (1992), "Can the Lombard effect be used to improve low voice intensity in Parkinson's disease?," *International Journal of Language & Communication Disorders*, Vol. 27 No. 2, pp. 121–127.

Arantes, P. (2015), "Time-normalization of fundamental frequency contours: A hands-on tutorial," *Courses on Speech Prosody*, No. September.

Bapineedu, G. (2010), *Analysis of Lombard Effect Speech and Its Application in Speaker Verification for Imposter Detection*, Ph.D. dissertation, Language Technologies Research Centre, International Institute of Information Technology (M.Sc. thesis).

Barber, D. (2011), *Bayesian Reasoning and Machine Learning*, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, Cambridge, available at:https://doi.org/10.1017/CBO9780511804779.

Beerends, J.G., van Buuren, R., van Vugt, J. and Verhave, J. (2009), "Objective speech intelligibility measurement on the basis of natural speech in combination with perceptual modeling," *AES: Journal of the Audio Engineering Society*, Vol. 57 No. 5, pp. 299–308.

Beerends, J.G., Hekstra, A.P., Rix, A.W. and Hollier, M.P. (2002), "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. Part II - Psychoacoustic model," *AES: Journal of the Audio Engineering Society*, Vol. 50 No. 10.

Beerends, J.G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J. and Keyhl, M. (2013), "Perceptual Objective Listening Quality Assessment (POLQA), the third generation ITU-T standard for end-to-end speech Quality measurement Part II-perceptual model," *AES: Journal of the Audio Engineering Society*, Vol. 61 No. 6.

Benesty, J., Sondhi, M. M., Yiteng, A.H. (2008), "Handbook of Speech Processing", Springer Handbooks; available at: (https://link.springer.com/content/pdf/10.1007/978-3-540-49127-9.pdf)

Bhavan, A., Chauhan, P., Hitkul and Shah, R.R. (2019), "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, Vol. 184, available at:https://doi.org/10.1016/j.knosys.2019.104886.

Bishop, Ch. M. (2006), "Pattern Recognition and Machine Learning, Springer-Verlag New York; available at: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20%20Pattern%20Recognitio%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf (accessed May 2022)

Boersma, P. and Weenink, D. (2018), "Praat: doing phonetics by computer [Computer program]. Version 6.0.43," *Retrieved 8 September 2018*.

Bollepalli, B., Juvela, L., Airaksinen, M., Valentini-Botinhao, C. and Alku, P. (2019), "Normal-to-Lombard adaptation of speech synthesis using long short-term memory recurrent neural networks," *Speech Communication*, Vol. 110, available at:https://doi.org/10.1016/j.specom.2019.04.008.

Bořil, H., Fousek, P. and Höge, H. (2007), "Two-stage system for robust neutral/lombard speech recognition," *Interspeech 2007*, Vol. 4, ISCA, ISCA, pp. 1074–1077.

Bořil, H., Fousek, P., Sündermann, D., Červa, P. and Žďánský, J. (2006), "Lombard Speech Recognition: A Comparative Study," *Proc. 16th Czech-German Workshop on Speech Processing*.

Brumm, H. and Zollinger, S.A. (2011), "The evolution of the Lombard effect: 100 years of psychoacoustic research," *Behaviour*, Vol. 148 No. 11–13, pp. 1173–1198.

Byrd, R.H., Lu, P., Nocedal, J. and Zhu, C. (1995), "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, Vol. 16 No. 5, pp. 1190–1208.

Cooke, M., Aubanel, V. and García Lecumberri, M.L. (2019), "Combining spectral and temporal modification techniques for speech intelligibility enhancement," *Computer Speech and Language*, Vol. 55, pp. 26–39.

Cooper, E. and Hirschberg, J. (2018), "Adaptation and frontend features to improve naturalness in found-data synthesis," *Proceedings of the International Conference on Speech Prosody*, Vol. 2018-June, available at:https://doi.org/10.21437/SpeechProsody.2018-160.

Corretge, R. (2012), "Praat Vocal Toolkit," available at: http://www.praatvocaltoolkit.com (accessed 14 May 2022).

Cortes, C., Haffner, P. and Mohri, M. (2004), "Rational kernels: Theory and algorithms," *Journal of Machine Learning Research*, Vol. 5, pp. 1035–1062.

Cortes, C. and Vapnik, V. (1995), "Support-Vector Networks," *Machine Learning*, Vol. 20 No. 3, pp. 273–297.

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J. and Szykulski, M. (2017), "An audio-visual corpus for multimodal automatic speech recognition," *Journal of Intelligent Information Systems*, Vol. 49 No. 2, pp. 1–26.

Darwin, C.J., Brungart, D.S. and Simpson, B.D. (2003), "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *The Journal of the Acoustical Society of America*, Vol. 114 No. 5, pp. 2913–2922.

Deng, J., Xu, X., Zhang, Z., Fruhholz, S. and Schuller, B. (2016), "Exploitation of Phase-Based Features for Whispered Speech Emotion Recognition," *IEEE Access*, Vol. 4, pp. 4299–4309.

Dubey, R.K. and Kumar, A. (2015), "Comparison of subjective and objective speech quality assessment for different degradation / noise conditions," *2015 International Conference on Signal Processing and Communication, ICSC 2015*, pp. 261–266.

Dubnov, S. (2004), "Generalization of spectral flatness measure for non-Gaussian linear processes," *IEEE Signal Processing Letters*, Vol. 11 No. 8, available at:https://doi.org/10.1109/LSP.2004.831663.

Duch W., Korbicz J., Rutkowski L. and Tadeusiewicz R. (2000), „Sieci neuronowe" (*in Polish*), Akademicka Oficyna Wydawnicza EXIT; Warszawa 2000.

Egan, J.J. (1972), "Psychoacoustics of the Lombard voice response.," *Journal of Auditory Research*, Vol. 12 No. 4, pp. 318–324.

Ellis, D. (2002), "Aurora noise database," available at: https://www.ee.columbia.edu/~dpwe/sounds/noise/ (accessed 14 May 2022).

Ellis, D.P.W. (2003), "Sinewave and sinusoid+ noise analysis/synthesis in Matlab. online web resource," available at: http://www.ee.columbia.edu/dpwe/resources/matlab/sinemodel (accessed 1 November 2019).

Ellis, D.P.W. and Weiss, R.J. (2006), "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Vol. 5*, Vol. 5, pp. V–V.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014), "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, Vol. 15, pp. 3133–3181.

Gałka, J., Grzybowska, J., Igras, M., Jaciów, P., Wajda, K., Witkowski, M., Ziółko, M. (2015) "System supporting speaker identification in emergency call center," *Proc. Interspeech* 2015, 724-725.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), "Deep Learning", MIT Press, 2016; available at: https://www.deeplearningbook.org/.

Gosztolya, G. (2019), "Posterior-thresholding feature extraction for paralinguistic speech classification," *Knowledge-Based Systems*, Vol. 186, available at:https://doi.org/10.1016/j.knosys.2019.104943.

Ho, T.K. (1995), "Random Decision Forests," *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995*, pp. 278–282.

Hsu, J. (2021), "PyWORLD - A Python wrapper of WORLD Vocoder," available at: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder (accessed 16 May 2022).

ITU-T Recommendation BS.1534-1. (2003), "BS.1534: Method for the subjective assessment of intermediate quality level of audio systems," available at: https://www.itu.int/rec/R-REC-BS.1534-1-200301-S/en (accessed 14 May 2022).

ITU-T Recommendation P.563. (2004), "Single-ended method for objective speech quality assessment in narrow-band telephony applications," *ITU-T Recommendation P.563*.

ITU-T Recommendation P.800. (1996), "Recommendation ITU-T P.800 Methods for subjective determination of transmission quality," *International Telecommunication Union*, Vol. 800.

ITU-T Recommendation P.800.1. (2006), "Recommendation P. 800.1: Mean opinion score (MOS) terminology," *ITU-T P-Series, The International Telecommunication Union (ITU)*.

ITU-T Recommendation P.862. (2001), "P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," available at: https://www.itu.int/rec/T-REC-P.862-200102-I/en (accessed 14 May 2022).

ITU-T Recommendation P.862.1. (2003), "P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO," *ITU-T Recommendation*.

James, G. (n.d.). "UC Business Analytics R Programming Guide," available at: https://uc-r.github.io/discriminant_analysis (accessed 14 May 2022).

Jensen, J. and Taal, C.H. (2016), "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 24 No. 11, pp. 2009–2022.

Jokinen, E., Takanen, M., Vainio, M. and Alku, P. (2014), "An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech," *Computer Speech & Language*, Vol. 28 No. 2, pp. 619–628.

Junqua, J.-C., Fincke, S. and Field, K. (1999), "The Lombard effect: a reflex to better communicate with others in noise," *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, IEEE, pp. 2083–2086 vol.4.

Kamiński, B., Jakubczyk, M. and Szufel, P. (2018), "A framework for sensitivity analysis of decision trees," *Central European Journal of Operations Research*, Vol. 26 No. 1, pp. 135–159.

Kawahara, H. (2006), "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, Vol. 27 No. 6, pp. 349–353.

Kawahara, H., Cheveigné, A. de, Banno, H., Takahashi, T. and Irino, T. (2005), "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," *Interspeech 2005*, ISCA, ISCA, pp. 537–540.

Klapuri, A. and Davy, M. (2006), *Signal Processing Methods for Music Transcription*, edited by Klapuri, A. and Davy, M.*Signal Processing Methods for Music Transcription*, Springer Science and Business Media LLC., Boston, MA, available at: https://doi.org/10.1007/0-387-32845-9.

Kleczkowski, P., Zak, A. and Król-Nowak, A. (2017), "Lombard Effect in Polish Speech and its Comparison in English Speech," *Archives of Acoustics*, Vol. 42 No. 4, pp. 561–569.

Koszuta, S. and Szklanny, K. (2017), "Implementation and verification of speech database for unit selection speech synthesis, " in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, pp. 1263–1267, https://annals-csis.org/proceedings/2017/drp/pdf/395.pdf

Korvel, G., Kakol, K., Kurasova, O. and Kostek, B. (2020), "Evaluation of Lombard Speech Models in the Context of Speech in Noise Enhancement," *IEEE Access*, Vol. 8, pp. 155156–155170.

Korvel, G., Kurasova, O. and Kostek, B. (2019), "An Attempt to Create Speech Synthesis Model That Retains Lombard Effect Characteristics," *Proceedings of the 16th International Joint Conference on E-Business and Telecommunications*, Vol. 1, SCITEPRESS - Science and Technology Publications, pp. 280–289.

Korvel, G., Šimonytė, V. and Slivinskas, V. (2016), "A Phoneme Harmonic Generator," *Information Technology And Control*, Vol. 45 No. 1, pp. 7–12.

Kraljevski, I., Chungurski, S., Stojanovic, I. and Arsenovski, S. (2010), "Synthesized speech quality evaluation using ITU-T P.563," *18th Telecommunications Forum TELFOR*.

Krčadinac, O., Šošević, U. and Starčević, D. (2021), "Evaluating the Performance of Speaker Recognition Solutions in E-Commerce Applications," *Sensors*, Vol. 21 No. 18, p. 6231.

Kurban Ubul, Askar Hamdulla and Alim Aysa. (2009), "A Digital Signal Processing teaching methodology using Praat," *2009 4th International Conference on Computer Science & Education*, IEEE, pp. 1804–1809.

Laroche, J. and Dolson, M. (1999), "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, Vol. 7 No. 3, pp. 323–332.

Lau, P. (2008), "The Lombard Effect as a Communicative Phenomenon," *UC Berkeley Phonology Lab Annual Reports*, Vol. 4, available at:https://doi.org/10.5070/P719J8J0B6.

LeCun, Y., Haffner, P., Bottou, L. and Bengio, Y. (1999), "Object Recognition with Gradient-Based Learning," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 1681, pp. 319–345.

Lee, Y. K., Kim H. W., and Park, J. G. (2021), "Many-to-Many Unsupervised Speech Conversion From Nonparallel Corpora," in *IEEE Access*, vol. 9, pp. 27278-27286, doi: 10.1109/ACCESS.2021.3058382.

Li, J. (2021), "Recent Advances in End-to-End Automatic Speech Recognition," *Invited Paper Submitted to APSIPA Transactions on Signal and Information Processing*, available at: http://arxiv.org/abs/2111.01690.

Li, J., Deng, L., Haeb-Umbach, R. and Gong, Y. (2016), *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, Elsevier, available at:https://doi.org/10.1016/C2014-0-02251-4.

Lombard, E. (1911), "Le signe de l'élévation de la voix (translated from French)," *Annales Des Maladies de l'oreille et Du Larynx*, Vol. 37 No. 2, pp. 101–119.

López, A.R., Seshadri, S., Juvela, L., Räsänen, O. and Alku, P. (2017), "Speaking Style Conversion from Normal to Lombard Speech Using a Glottal Vocoder and Bayesian GMMs," *Interspeech 2017*, Vol. 2017-August, ISCA, ISCA, pp. 1363–1367.

Lu, Y. and Cooke, M. (2008), "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, Vol. 124 No. 5, pp. 3261–3275.

Lu, Y. and Cooke, M. (2009), "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, Vol. 51 No. 12, pp. 1253–1262.

Luong, M., Tran, V.A. (2021), "Many-to-Many Voice Conversion Based Feature Disentanglement Using Variational Autoencoder," *Proc. Interspeech* 2021, 851-855, doi: 10.21437/Interspeech.2021-2086.

Łatka, Z., Gałka, J. and Ziółko, B. (2019), "Cross-gender Voice Conversion with Constant F0-Ratio and Average Background Conversion Model, " *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6825-6829, doi: 10.1109/ICASSP.2019.8683369.

Makowski, R. and Hossa R. (2020), "Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise," *Applied Acoustics*, vol. 166, p. 107344, https://doi.org/10.1016/j.apacoust.2020.107344.

Marxer, R., Barker, J., Alghamdi, N. and Maddock, S. (2018), "The impact of the Lombard effect on audio and visual speech recognition systems," *Speech Communication*, Vol. 100, pp. 58–68.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E. and Nieto, O. (2015), "librosa: Audio and Music Signal Analysis in Python," *Proceedings of the 14th Python in Science Conference*, pp. 18–24.

Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M. and di Natale, C. (2014), "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure," *Knowledge-Based Systems*, Vol. 63, pp. 68–81.

Mermelstein, P. (1976), "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*.

Merritt, T., Ezzerg, A., Biliński, P., Proszoewska, M., Pokora, K., Barra-Chicote, R., Korzekwa, D., (2022), "Text-Free Non-Parallel Many-To-Many Voice Conversion Using Normalising Flow," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6782-6786, doi: 10.1109/ICASSP43922.2022.9746368.

Michelsanti, D., Tan, Z.-H., Sigurdsson, S. and Jensen, J. (2019), "Deep-learning-based audio-visual speech enhancement in presence of Lombard effect," *Speech Communication*, Vol. 115, pp. 38–50.

Morise, M. (2012), "PLATINUM: A method to extract excitation signals for voice synthesis system," *Acoustical Science and Technology*, Vol. 33 No. 2, pp. 123–125.

Morise, M. (2015), "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, Vol. 67, pp. 1–7.

Morise, M. (2016), "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, Vol. 84, pp. 57–65.

Morise, M. (2017), "Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals," *Interspeech 2017*, Vol. 2017-August, ISCA, ISCA, pp. 2321–2325.

Morise, M., Kawahara, H. and Katayose, H. (2009), "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," *Audio Engineering Society Conference: 35th International Conference: Audio for Game*.

Morise, M., Yokomori, F. and Ozawa, K. (2016), "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, Vol. E99.D No. 7, pp. 1877–1884.

Moulines, E. and Charpentier, F. (1990), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, Vol. 9 No. 5–6, pp. 453–467.

Mu, W., Yin, B., Huang, X., Xu, J. and Du, Z. (2021), "Environmental sound classification using temporal-frequency attention based convolutional neural network," *Scientific Reports*, Vol. 11 No. 1, p. 21552.

Müller, M. (2015), *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, *Springer*, Springer.

Odya, P., Kotus, J., Kurowski, A. and Kostek, B. (2021), "Acoustic Sensing Analytics Applied to Speech in Reverberation Conditions," *Sensors*, Vol. 21 No. 18, p. 6320.

O'Shaughnessy, D. (1988), "Linear predictive coding," *IEEE Potentials*, Vol. 7 No. 1, pp. 29–32.

Pan, S.J. and Yang, Q. (2010), "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, available at:https://doi.org/10.1109/TKDE.2009.191.

Patel, R. and Schell, K.W. (2008), "The Influence of Linguistic Content on the Lombard Effect," *Journal of Speech, Language, and Hearing Research*, Vol. 51 No. 1, pp. 209–220.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., *et al.* (2011), "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, Vol. 12.

Pick, H.L., Siegel, G.M., Fox, P.W., Garber, S.R. and Kearney, J.K. (1989), "Inhibiting the Lombard effect," *The Journal of the Acoustical Society of America*, Vol. 85 No. 2, pp. 894–900.

Platt, J. (1999), "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers, MIT Press*, Vol. 10 No. 3.

Pyž, G., Šimonytė, V. and Slivinskas, V. (2014), "Developing Models of Lithuanian Speech Vowels and Semivowels," *Informatica*, Vol. 25 No. 1, pp. 55–72.

Raitio, T., Suni, A., Vainio, M. and Alku, P. (2011), "Analysis of HMM-based Lombard speech synthesis," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, available at:https://doi.org/10.21437/interspeech.2011-696.

Rasmussen, C.E. and Williams, C.K.I. (2006), *Gaussian Processes for Machine Learning. 2006*, *The MIT Press, Cambridge, MA, USA*, MIT Press., Vol. 38.

Rix, A.W., Hollier, M.P., Hekstra, A.P. and Beerends, J.G. (2002), "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. Part I - Time-delay compensation," *AES: Journal of the Audio Engineering Society*, Vol. 50 No. 10.

Rojas, R. (2009), "AdaBoost and the Super Bowl of Classifiers A Tutorial Introduction to Adaptive Boosting," *Writing*.

Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R. and Zurada, J.M. (Eds.) (2021), "Artificial Intelligence and Soft Computing", 20th International Conference ICAISC 2021, Proceedings, Part I, LNCS 12854, Springer-Verlag, Berlin – Heidelberg – New York, available at: doi: https://doi.org/10.1007/978-3-030-87986-0.

Saba, J.N., and Hansen, J.H.L. (2022), "The effects of Lombard perturbation on speech intelligibility in noise for normal hearing and cochlear implant listeners," *The Journal of the Acoustical Society of America*, 151, 1007; https://doi.org/10.1121/10.0009377.

Sagisaka, Y. (1988), "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," ICASSP-88., *International Conference on Acoustics, Speech, and Signal Processing*, IEEE, pp. 679–682.

Seshadri, S., Juvela, L., Rasanen, O. and Alku, P. (2019), "Vocal Effort Based Speaking Style Conversion Using Vocoder Features and Parallel Learning," *IEEE Access*, Vol. 7, pp. 17230–17246.

Soloducha, M., Raake, A., Kettler, F. and Voigt, P. (2016), "Lombard speech database for German language Recording setup," *Proc. DAGA*.

Stathopoulos, E.T., Huber, J.E., Richardson, K., Kamphaus, J., DeCicco, D., Darling, M., Fulcher, K., *et al.* (2014), "Increased vocal intensity due to the Lombard effect in speakers with

Parkinson's disease: Simultaneous laryngeal and respiratory strategies," *Journal of Communication Disorders*, Vol. 48 No. 1, pp. 1–17.

Statista. (2022), "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025," *Statista Research Department*, available at: https://www.statista.com/statistics/871513/worldwide-data-created/ (accessed 14 April 2022).

Stowe, L.M. and Golob, E.J. (2013), "Evidence that the Lombard effect is frequency-specific in humans," *The Journal of the Acoustical Society of America*, Vol. 134 No. 1, pp. 640–647.

Tabachnick, B.G. and Fidell, L.S. (2007), *Experimental Designs Using ANOVA*, *Experimental Design Using Anova*, Thomson/Brooks/Cole, Belmont, CA.

Tadeusiewicz, R. (1988), „Sygnał mowy" (*in Polish*), Warszawa WKiŁ.

Takaki, S., Kim, S. and Yamagishi, J. (2016), "Speaker Adaptation of Various Components in Deep Neural Network based Speech Synthesis," *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, ISCA, ISCA, pp. 153–159.

Therrien, A.S., Lyons, J. and Balasubramaniam, R. (2012), "Sensory Attenuation of Self-Produced Feedback: The Lombard Effect Revisited," edited by Gribble, P.L.*PLoS ONE*, Vol. 7 No. 11, p. e49370.

Tuncer, T., Dogan, S. and Acharya, U.R. (2021), "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowledge-Based Systems*, Vol. 211, p. 106547.

Vlaj, D. and Kacic, Z. (2011), "The Influence of Lombard Effect on Speech Recognition," *Speech Technologies, Chapter 7*, InTech, pp. 151–168.

Wang, Y., and Nishizaki, H. (2022), "Combination of Time-domain, Frequency-domain, and Cepstral-domain Acoustic Features for Speech Commands Classification," https://arxiv.org/pdf/2203.16085.pdf.

Watanabe, S., Delcroix, M., Metze, F. and Hershey, J.R. (2017), *New Era for Robust Speech Recognition*, edited by Watanabe, S., Delcroix, M., Metze, F. and Hershey, J.R.*New Era for Robust Speech Recognition*, Springer International Publishing, Cham, available at:https://doi.org/10.1007/978-3-319-64680-0.

webMUSHRA. (2019), "webMUSHRA," available at: http://www.ee.columbia.edu/dpwe/resources/matlab/sinemodel/ (accessed 1 November 2021).

Wu, T.F., Lin, C.J. and Weng, R.C. (2004), "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, Vol. 5, pp. 975–1005.

Zalkow, F. and Müller, M. (n.d.). "Log-Frequency Spectrogram and Chromagram," available at: https://www.audiolabs-erlangen.de/resources/MIR/FMP/C3/C3S1_SpecLogFreq-Chromagram.html (accessed 14 May 2022).

Zhang, H. (2004), "The optimality of Naive Bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, Vol. 2.

Zhu, C., Byrd, R.H., Lu, P. and Nocedal, J. (1997), "Algorithm 778: L-BFGS-B," *ACM Transactions on Mathematical Software*, Vol. 23 No. 4, pp. 550–560.

Zielinski, S. (2016), "On Some Biases Encountered in Modern Audio Quality Listening Tests (Part 2): Selected Graphical Examples and Discussion," *Journal of the Audio Engineering Society*, Vol. 64 No. 1/2, pp. 55–74.

Zollinger, S.A. and Brumm, H. (2011), "The Lombard effect," *Current Biology*, Vol. 21 No. 16, pp. R614–R615.

# Appendix A Detailed results of preliminary experiments

### Finding the best modification method for the speech signal in presence of noise

Average values of the P.563 metrics for all recordings after modifications are presented in Figure A.1. Figures A.2 and A.3 present the results divided by gender to observe the difference between the speech modifications' results for female and male voices. The results are averaged for all noise types.

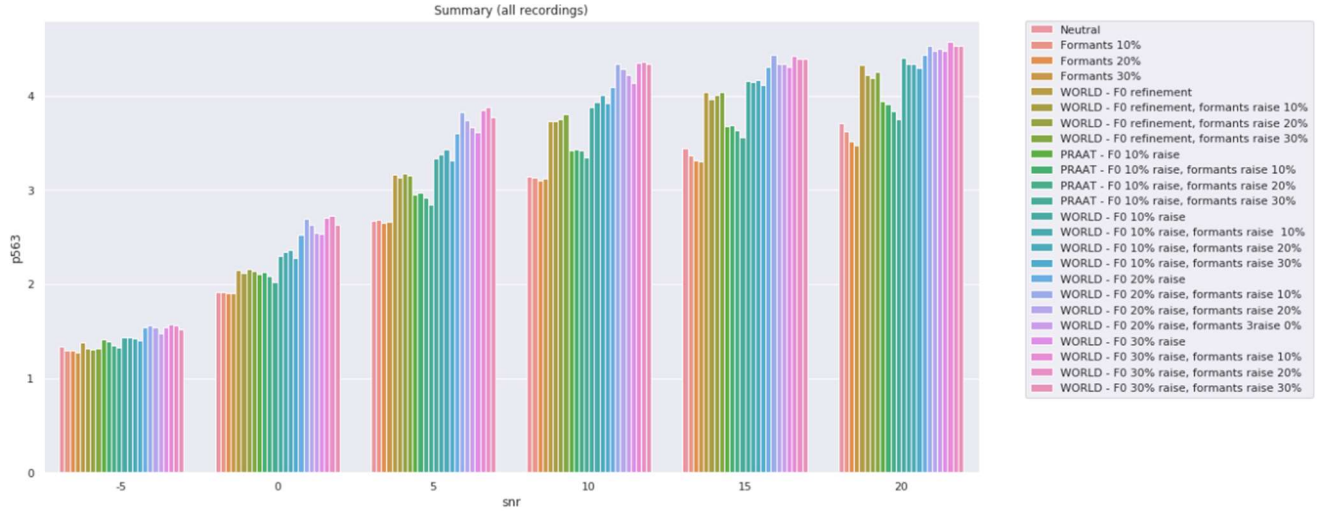



*Figure A.1 Averaged results of P.563 for all recordings, neutral and modified, by signal-to-noise ratio.*

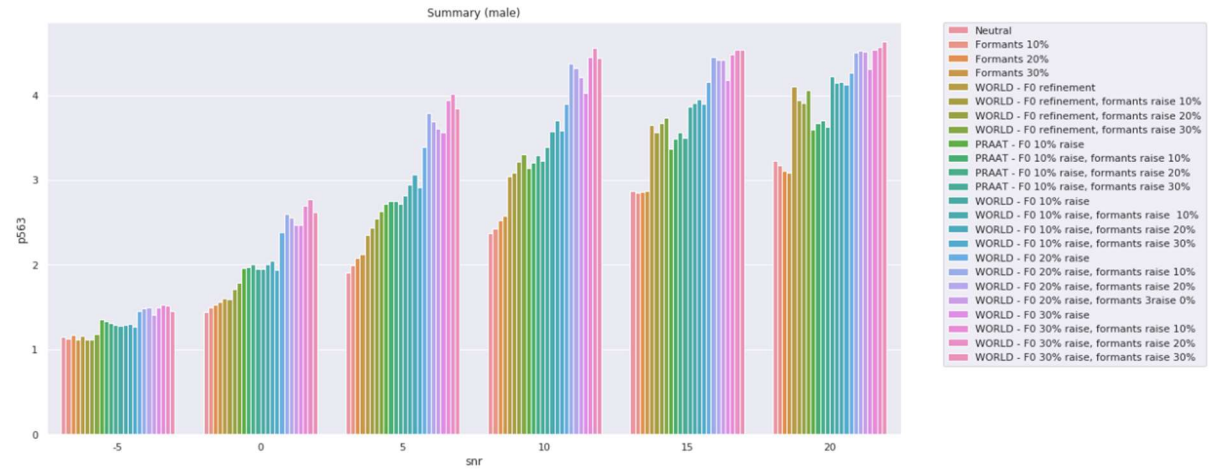



*Figure A.2 Averaged results of P.563 for male recordings, neutral and modified, by signal-to-noise ratio.*

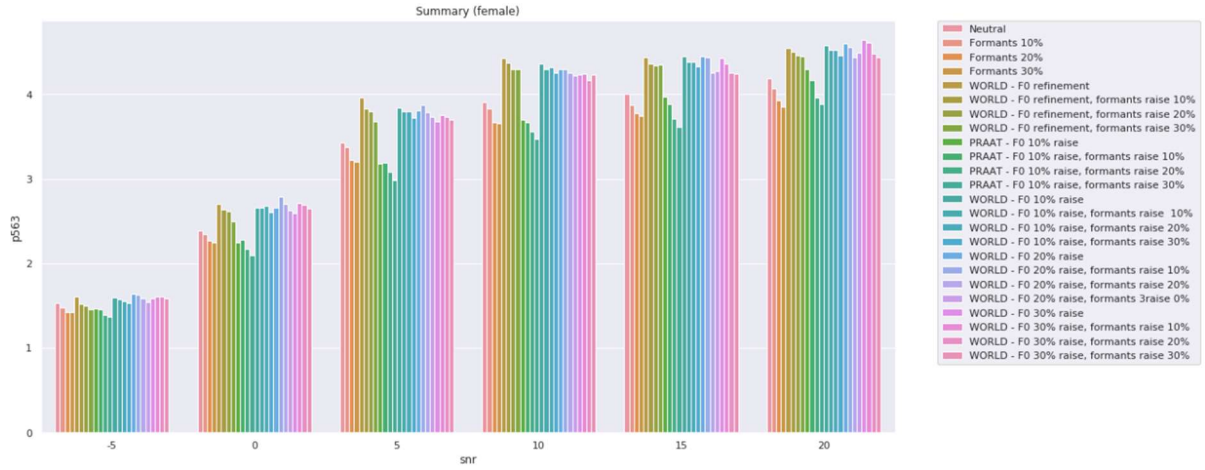



*Figure A.3 Averaged results of P.563 for female recordings, neutral and modified, by signal-to-noise ratio.*

Averaged results are presented in Tables A.1-A.12. The best results are highlighted in bold, red font.

*Table A.1 Average P.563 results for all recordings*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 1.30 | 1.92 | 2.68 | 3.13 | 3.36 | 3.62 |
| Formants 20% | | 1.30 | 1.90 | 2.65 | 3.10 | 3.32 | 3.52 |
| Formants 30% | | 1.28 | 1.91 | 2.67 | 3.12 | 3.31 | 3.47 |
| Neutral | | 1.34 | 1.92 | 2.67 | 3.14 | 3.44 | 3.71 |
| PRAAT - F0 10% raise | | 1.41 | 2.11 | 2.95 | 3.42 | 3.67 | 3.94 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.39 | 2.13 | 2.97 | 3.43 | 3.69 | 3.91 |
| PRAAT - F0 10% raise, formants raise 20% | | 1,35 | 2,09 | 2,92 | 3,42 | 3,63 | 3,83 |
| PRAAT - F0 10% raise, formants raise 30% | | 1,32 | 2,02 | 2,85 | 3,35 | 3,56 | 3,75 |
| WORLD - F0 10% raise | | 1.44 | 2.30 | 3.33 | 3.88 | 4.16 | 4.40 |
| WORLD - F0 10% raise, formants raise 10% | | 1.44 | 2.34 | 3.38 | 3.93 | 4.14 | 4.34 |
| WORLD - F0 10% raise, formants raise 20% | | 1.43 | 2.36 | 3.43 | 4.01 | 4.17 | 4.34 |
| WORLD - F0 10% raise, formants raise 30% | | 1.40 | 2.28 | 3.32 | 3.92 | 4.11 | 4.30 |
| WORLD - F0 20% raise | | 1.55 | 2.52 | 3.60 | 4.10 | 4.30 | 4.43 |
| WORLD - F0 20% raise, formants 3raise 0% | | 1.48 | 2.55 | 3.67 | 4.22 | 4.34 | 4.50 |
| WORLD - F0 20% raise, formants raise 10% | | 1.56 | 2.69 | 3.83 | 4.33 | 4.44 | 4.53 |
| WORLD - F0 20% raise, formants raise 20% | | 1.54 | 2.62 | 3.74 | 4.29 | 4.34 | 4.48 |
| WORLD - F0 30% raise | | 1.54 | 2.53 | 3.62 | 4.13 | 4.30 | 4.48 |
| WORLD - F0 30% raise, formants raise 10% | | **1.57** | 2.71 | 3.85 | 4.35 | **4.42** | **4.57** |
| WORLD - F0 30% raise, formants raise 20% | | 1.56 | **2.73** | **3.88** | **4.36** | 4.39 | 4.53 |
| WORLD - F0 30% raise, formants raise 30% | | 1.52 | 2.63 | 3.77 | 4.33 | 4.39 | 4.53 |
| WORLD - F0 refinement | | 1.38 | 2.15 | 3.16 | 3.73 | 4.04 | 4.33 |
| WORLD - F0 refinement, formants raise 10% | | 1.32 | 2.12 | 3.13 | 3.73 | 3.96 | 4.22 |
| WORLD - F0 refinement, formants raise 20% | | 1.31 | 2.16 | 3.17 | 3.75 | 4.00 | 4.19 |
| WORLD - F0 refinement, formants raise 30% | | 1.32 | 2.14 | 3.15 | 3.80 | 4.04 | 4.25 |

*Table A.2 Average P.563 results for male voice recordings*

| Type of modification | SNR: | P.563 -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
|---|---|---|---|---|---|---|---|
| Formants 10% | | 1.13 | 1.49 | 2.00 | 2.43 | 2.85 | 3.18 |
| Formants 20% | | 1.17 | 1.53 | 2.08 | 2.52 | 2.86 | 3.11 |
| Formants 30% | | 1.12 | 1.56 | 2.13 | 2.57 | 2.87 | 3.09 |
| Neutral | | 1.15 | 1.44 | 1.91 | 2.37 | 2.87 | 3.23 |
| PRAAT - F0 10% raise | | 1.35 | 1.97 | 2.72 | 3.14 | 3.37 | 3.59 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.33 | 1.98 | 2.76 | 3.20 | 3.49 | 3.67 |
| PRAAT - F0 10% raise, formants raise 20% | | 1.31 | 2.00 | 2.75 | 3.29 | 3.56 | 3.70 |
| PRAAT - F0 10% raise, formants raise 30% | | 1.28 | 1.95 | 2.71 | 3.23 | 3.50 | 3.62 |
| WORLD - F0 10% raise | | 1.28 | 1.95 | 2.82 | 3.39 | 3.86 | 4.22 |
| WORLD - F0 10% raise, formants raise 10% | | 1.30 | 2.01 | 2.95 | 3.57 | 3.91 | 4.15 |
| WORLD - F0 10% raise, formants raise 20% | | 1.30 | 2.04 | 3.06 | 3.70 | 3.95 | 4.16 |
| WORLD - F0 10% raise, formants raise 30% | | 1.26 | 1.94 | 2.92 | 3.58 | 3.89 | 4.13 |
| WORLD - F0 20% raise | | 1.45 | 2.38 | 3.39 | 3.89 | 4.16 | 4.27 |
| WORLD - F0 20% raise, formants 3raise 0% | | 1.41 | 2.47 | 3.61 | 4.21 | 4.41 | 4.51 |
| WORLD - F0 20% raise, formants raise 10% | | 1.49 | 2.59 | 3.79 | 4.37 | 4.44 | 4.51 |
| WORLD - F0 20% raise, formants raise 20% | | 1.50 | 2.55 | 3.69 | 4.32 | 4.42 | 4.52 |
| WORLD - F0 30% raise | | 1.49 | 2.47 | 3.56 | 4.03 | 4.18 | 4.31 |
| WORLD - F0 30% raise, formants raise 10% | | **1.53** | 2.70 | 3.94 | 4.45 | 4.48 | 4.54 |
| WORLD - F0 30% raise, formants raise 20% | | 1.52 | **2.77** | **4.02** | **4.55** | **4.54** | 4.57 |
| WORLD - F0 30% raise, formants raise 30% | | 1.45 | 2.62 | 3.85 | 4.44 | **4.54** | **4.63** |
| WORLD - F0 refinement | | 1.16 | 1.60 | 2.36 | 3.04 | 3.65 | 4.10 |
| WORLD - F0 refinement, formants raise 10% | | 1.12 | 1.60 | 2.43 | 3.09 | 3.57 | 3.94 |
| WORLD - F0 refinement, formants raise 20% | | 1.11 | 1.71 | 2.55 | 3.21 | 3.67 | 3.91 |
| WORLD - F0 refinement, formants raise 30% | | 1.18 | 1.78 | 2.63 | 3.30 | 3.74 | 4.05 |

*Table A.3 Average P.563 results for female voice recordings*

| Type of modification | SNR: | P.563 -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
|---|---|---|---|---|---|---|---|
| Formants 10% | | 1.47 | 2.34 | 3.37 | 3.83 | 3.88 | 4.07 |
| Formants 20% | | 1.42 | 2.27 | 3.22 | 3.67 | 3.77 | 3.93 |
| Formants 30% | | 1.43 | 2.25 | 3.21 | 3.66 | 3.74 | 3.85 |
| Neutral | | 1.53 | 2.39 | 3.43 | 3.90 | 4.01 | 4.19 |
| PRAAT - F0 10% raise | | 1.47 | 2.25 | 3.18 | 3.70 | 3.97 | 4.29 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.45 | 2.28 | 3.19 | 3.67 | 3.88 | 4.16 |
| PRAAT - F0 10% raise, formants raise 20% | | 1.39 | 2.17 | 3.08 | 3.55 | 3.71 | 3.96 |
| PRAAT - F0 10% raise, formants raise 30% | | 1.37 | 2.10 | 2.98 | 3.47 | 3.62 | 3.88 |
| WORLD - F0 10% raise | | 1.59 | 2.66 | 3.84 | 4.36 | **4.45** | 4.58 |
| WORLD - F0 10% raise, formants raise 10% | | 1.58 | 2.66 | 3.80 | 4.30 | 4.38 | 4.53 |
| WORLD - F0 10% raise, formants raise 20% | | 1.55 | 2.68 | 3.80 | 4.32 | 4.38 | 4.53 |
| WORLD - F0 10% raise, formants raise 30% | | 1.53 | 2.61 | 3.72 | 4.25 | 4.33 | 4.46 |

| | | | | | | |
|---|---|---|---|---|---|---|
| WORLD - F0 20% raise | **1.64** | 2.66 | 3.81 | 4.30 | 4.44 | **4.60** |
| WORLD - F0 20% raise, formants 3raise 0% | 1.54 | 2.62 | 3.73 | 4.22 | 4.27 | 4.49 |
| WORLD - F0 20% raise, formants raise 10% | 1.63 | **2.79** | **3.87** | 4.29 | 4.44 | 4.55 |
| WORLD - F0 20% raise, formants raise 20% | 1.58 | 2.70 | 3.79 | 4.25 | 4.25 | 4.44 |
| WORLD - F0 30% raise | 1.58 | 2.59 | 3.68 | 4.23 | 4.43 | 4.64 |
| WORLD - F0 30% raise, formants raise 10% | 1.61 | 2.71 | 3.76 | 4.24 | 4.36 | **4.60** |
| WORLD - F0 30% raise, formants raise 20% | 1.61 | 2.69 | 3.74 | 4.17 | 4.25 | 4.48 |
| WORLD - F0 30% raise, formants raise 30% | 1.58 | 2.65 | 3.70 | 4.23 | 4.25 | 4.43 |
| WORLD - F0 refinement | 1.61 | 2.70 | 3.96 | **4.43** | 4.44 | 4.55 |
| WORLD - F0 refinement, formants raise 10% | 1.52 | 2.64 | 3.83 | 4.37 | 4.36 | 4.50 |
| WORLD - F0 refinement, formants raise 20% | 1.50 | 2.62 | 3.79 | 4.30 | 4.34 | 4.46 |
| WORLD - F0 refinement, formants raise 30% | 1.45 | 2.50 | 3.68 | 4.30 | 4.35 | 4.45 |

*Table A.4 Average P.563 results for recordings with "airport" noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 1.15 | 2.06 | 2.81 | 3.19 | 3.48 | 3.65 |
| Formants 20% | | 1.12 | 1.95 | 2.76 | 3.17 | 3.31 | 3.52 |
| Formants 30% | | 1.11 | 2.00 | 2.82 | 3.15 | 3.35 | 3.54 |
| Neutral | | 1.24 | 2.04 | 2.81 | 3.11 | 3.46 | 3.83 |
| PRAAT - F0 10% raise | | 1.20 | 2.20 | 3.12 | 3.43 | 3.76 | 4.03 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.20 | 2.26 | 3.13 | 3.51 | 3.74 | 3.96 |
| PRAAT - F0 10% raise, formants raise 20% | | 1.16 | 2.18 | 3.05 | 3.39 | 3.63 | 3.87 |
| PRAAT - F0 10% raise, formants raise 30% | | 1.12 | 2.09 | 2.94 | 3.37 | 3.50 | 3.82 |
| WORLD - F0 10% raise | | 1.35 | 2.55 | 3.60 | 3.91 | 3.92 | 4.55 |
| WORLD - F0 10% raise, formants raise 10% | | 1.30 | 2.51 | 3.59 | 3.98 | 4.25 | 4.43 |
| WORLD - F0 10% raise, formants raise 20% | | 1.24 | 2.56 | 3.66 | 4.11 | 4.20 | 4.39 |
| WORLD - F0 10% raise, formants raise 30% | | 1.24 | 2.40 | 3.58 | 4.07 | 4.16 | 4.35 |
| WORLD - F0 20% raise | | 1.37 | 2.77 | 3.88 | 4.03 | 4.22 | 4.51 |
| WORLD - F0 20% raise, formants 3raise 0% | | 1.26 | 2.71 | 3.87 | 4.38 | 4.34 | 4.53 |
| WORLD - F0 20% raise, formants raise 10% | | 1.36 | 2.99 | 4.09 | 4.36 | 4.41 | 4.54 |
| WORLD - F0 20% raise, formants raise 20% | | 1.32 | 2.86 | 4.01 | 4.34 | 4.35 | 4.49 |
| WORLD - F0 30% raise | | 1.35 | 2.75 | 3.88 | 4.11 | 4.16 | 4.51 |
| WORLD - F0 30% raise, formants raise 10% | | **1.41** | 2.93 | **4.16** | 4.43 | **4.44** | **4.61** |
| WORLD - F0 30% raise, formants raise 20% | | 1.35 | **2.94** | 4.14 | 4.40 | 4.40 | 4.52 |
| WORLD - F0 30% raise, formants raise 30% | | 1.37 | 2.82 | 4.02 | **4.45** | 4.38 | 4.53 |
| WORLD - F0 refinement | | 1.27 | 2.31 | 3.39 | 3.81 | 3.80 | 4.47 |
| WORLD - F0 refinement, formants raise 10% | | 1.14 | 2.31 | 3.43 | 3.81 | 4.01 | 4.28 |
| WORLD - F0 refinement, formants raise 20% | | 1.18 | 2.27 | 3.44 | 3.87 | 4.08 | 4.21 |
| WORLD - F0 refinement, formants raise 30% | | 1.17 | 2.23 | 3.33 | 3.88 | 4.06 | 4.26 |

*Table A.5 Average P.563 results for recordings with "babble speech" noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |

| | | | | | | |
|---|---|---|---|---|---|---|
| Formants 10% | 1.21 | 2.09 | 3.01 | 3.33 | 3.42 | 3.64 |
| Formants 20% | 1.20 | 2.03 | 2.97 | 3.26 | 3.40 | 3.61 |
| Formants 30% | 1.17 | 2.03 | 2.98 | 3.31 | 3.33 | 3.48 |
| Neutral | 1.24 | 2.13 | 2.94 | 3.33 | 3.51 | 3.82 |
| PRAAT - F0 10% raise | 1.24 | 2.31 | 3.25 | 3.63 | 3.80 | 3.98 |
| PRAAT - F0 10% raise, formants raise 10% | 1.25 | 2.36 | 3.23 | 3.62 | 3.77 | 3.90 |
| PRAAT - F0 10% raise, formants raise 20% | 1.27 | 2.28 | 3.23 | 3.67 | 3.68 | 3.79 |
| PRAAT - F0 10% raise, formants raise 30% | 1.22 | 2.15 | 3.16 | 3.47 | 3.60 | 3.72 |
| WORLD - F0 10% raise | 1.38 | 2.66 | 3.70 | 4.26 | 4.32 | 4.49 |
| WORLD - F0 10% raise, formants raise 10% | 1.41 | 2.68 | 3.74 | 4.21 | 4.25 | 4.36 |
| WORLD - F0 10% raise, formants raise 20% | 1.38 | 2.65 | 3.82 | 4.27 | 4.25 | 4.36 |
| WORLD - F0 10% raise, formants raise 30% | 1.29 | 2.63 | 3.70 | 4.16 | 4.25 | 4.32 |
| WORLD - F0 20% raise | 1.52 | 2.89 | 3.97 | 4.36 | 4.41 | 4.52 |
| WORLD - F0 20% raise, formants 3raise 0% | 1.43 | 2.88 | 4.04 | 4.49 | 4.39 | 4.52 |
| WORLD - F0 20% raise, formants raise 10% | 1.51 | 3.06 | 4.18 | 4.60 | **4.57** | 4.53 |
| WORLD - F0 20% raise, formants raise 20% | 1.52 | 3.00 | 4.10 | 4.57 | 4.44 | 4.50 |
| WORLD - F0 30% raise | 1.48 | 2.87 | 3.98 | 4.45 | 4.41 | 4.57 |
| WORLD - F0 30% raise, formants raise 10% | 1.57 | 3.12 | 4.22 | 4.63 | 4.54 | **4.59** |
| WORLD - F0 30% raise, formants raise 20% | **1.61** | **3.21** | **4.27** | **4.67** | 4.47 | 4.55 |
| WORLD - F0 30% raise, formants raise 30% | 1.51 | 3.02 | 4.14 | 4.64 | 4.50 | 4.56 |
| WORLD - F0 refinement | 1.35 | 2.50 | 3.55 | 4.09 | 4.25 | 4.47 |
| WORLD - F0 refinement, formants raise 10% | 1.29 | 2.42 | 3.54 | 3.98 | 4.08 | 4.20 |
| WORLD - F0 refinement, formants raise 20% | 1.29 | 2.42 | 3.51 | 4.03 | 4.15 | 4.24 |
| WORLD - F0 refinement, formants raise 30% | 1.25 | 2.48 | 3.54 | 4.10 | 4.15 | 4.30 |

*Table A.6 Average P.563 results for recordings with "train" noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 1.77 | 2.57 | 3.03 | 3.20 | 3.59 | 3.77 |
| Formants 20% | | 1.81 | 2.59 | 2.96 | 3.13 | 3.50 | 3.67 |
| Formants 30% | | 1.74 | 2.58 | 3.09 | 3.13 | 3.46 | 3.59 |
| Neutral | | 1.75 | 2.55 | 3.05 | 3.26 | 3.66 | 4.00 |
| PRAAT - F0 10% raise | | 1.89 | 2.75 | 3.32 | 3.57 | 3.83 | 4.15 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.94 | 2.81 | 3.40 | 3.54 | 3.89 | 4.12 |
| PRAAT - F0 10% raise, formants raise 20% | | 1.83 | 2.77 | 3.35 | 3.46 | 3.76 | 3.94 |
| PRAAT - F0 10% raise, formants raise 30% | | 1.83 | 2.69 | 3.24 | 3.40 | 3.73 | 3.87 |
| WORLD - F0 10% raise | | 2.06 | 3.11 | 3.88 | 3.76 | 4.49 | 4.64 |
| WORLD - F0 10% raise, formants raise 10% | | 2.04 | 3.15 | 3.91 | 4.00 | 4.31 | 4.53 |
| WORLD - F0 10% raise, formants raise 20% | | 2.04 | 3.25 | 3.94 | 3.99 | 4.31 | 4.46 |
| WORLD - F0 10% raise, formants raise 30% | | 1.91 | 3.04 | 3.81 | 3.92 | 4.29 | 4.40 |
| WORLD - F0 20% raise | | 2.26 | 3.40 | 4.16 | 4.00 | 4.43 | 4.62 |
| WORLD - F0 20% raise, formants 3raise 0% | | 2.23 | 3.50 | 4.28 | 4.12 | 4.44 | 4.57 |
| WORLD - F0 20% raise, formants raise 10% | | 2.36 | **3.75** | 4.40 | **4.30** | 4.49 | 4.63 |
| WORLD - F0 20% raise, formants raise 20% | | 2.27 | 3.59 | 4.31 | 4.25 | 4.44 | 4.54 |
| WORLD - F0 30% raise | | 2.30 | 3.46 | 4.21 | 4.05 | 4.44 | 4.66 |
| WORLD - F0 30% raise, formants raise 10% | | **2.46** | 3.72 | 4.39 | 4.28 | **4.56** | **4.67** |
| WORLD - F0 30% raise, formants raise 20% | | 2.40 | 3.73 | **4.41** | 4.27 | 4.45 | 4.58 |
| WORLD - F0 30% raise, formants raise 30% | | 2.26 | 3.62 | 4.33 | 4.21 | 4.44 | 4.59 |
| WORLD - F0 refinement | | 1.88 | 2.90 | 3.66 | 3.64 | 4.41 | 4.61 |

| | | | | | | |
|---|---|---|---|---|---|---|
| WORLD - F0 refinement, formants raise 10% | 1.83 | 2.85 | 3.61 | 3.90 | 4.22 | 4.40 |
| WORLD - F0 refinement, formants raise 20% | 1.80 | 2.93 | 3.65 | 3.90 | 4.20 | 4.35 |
| WORLD - F0 refinement, formants raise 30% | 1.84 | 2.92 | 3.65 | 3.92 | 4.26 | 4.42 |

*Table A.7 Average P.563 results for recordings with "subway" noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 1.01 | 1.44 | 2.53 | 3.00 | 3.23 | 3.53 |
| Formants 20% | | 0.93 | 1.46 | 2.50 | 3.03 | 3.28 | 3.49 |
| Formants 30% | | 0.94 | 1.52 | 2.46 | 2.95 | 3.20 | 3.39 |
| Neutral | | 1.07 | 1.43 | 2.48 | 3.01 | 3.34 | 3.68 |
| PRAAT - F0 10% raise | | 1.07 | 1.73 | 2.88 | 3.38 | 3.64 | 3.88 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.07 | 1.73 | 2.83 | 3.38 | 3.69 | 3.93 |
| PRAAT - F0 10% raise, formants raise 20% | | 1.03 | 1.69 | 2.75 | 3.44 | 3.66 | 3.84 |
| PRAAT - F0 10% raise, formants raise 30% | | 1.02 | 1.70 | 2.76 | 3.37 | 3.56 | 3.77 |
| WORLD - F0 10% raise | | 1.08 | 1.93 | 3.13 | 3.81 | 4.20 | 4.41 |
| WORLD - F0 10% raise, formants raise 10% | | 1.03 | 1.94 | 3.17 | 3.90 | 4.10 | 4.30 |
| WORLD - F0 10% raise, formants raise 20% | | 1.02 | 2.03 | 3.37 | 3.95 | 4.07 | 4.27 |
| WORLD - F0 10% raise, formants raise 30% | | 1.05 | 1.91 | 3.23 | 3.94 | 4.13 | 4.29 |
| WORLD - F0 20% raise | | 1.16 | 2.24 | 3.48 | 4.10 | 4.33 | 4.45 |
| WORLD - F0 20% raise, formants 3raise 0% | | 1.04 | 2.23 | 3.58 | 4.22 | 4.39 | 4.50 |
| WORLD - F0 20% raise, formants raise 10% | | 1.08 | 2.26 | 3.69 | 4.27 | 4.48 | 4.53 |
| WORLD - F0 20% raise, formants raise 20% | | **1.11** | 2.30 | 3.61 | 4.23 | 4.41 | 4.48 |
| WORLD - F0 30% raise | | 1.10 | 2.08 | 3.49 | 4.04 | 4.35 | 4.54 |
| WORLD - F0 30% raise, formants raise 10% | | 1.09 | 2.25 | 3.69 | 4.24 | 4.48 | **4.57** |
| WORLD - F0 30% raise, formants raise 20% | | 1.07 | **2.37** | **3.77** | **4.35** | **4.49** | 4.52 |
| WORLD - F0 30% raise, formants raise 30% | | 1.05 | 2.26 | 3.67 | 4.29 | 4.42 | 4.52 |
| WORLD - F0 refinement | | 1.08 | 1.73 | 3.00 | 3.62 | 4.13 | 4.37 |
| WORLD - F0 refinement, formants raise 10% | | 1.04 | 1.77 | 2.93 | 3.65 | 3.90 | 4.22 |
| WORLD - F0 refinement, formants raise 20% | | 0.97 | 1.79 | 3.05 | 3.72 | 3.96 | 4.11 |
| WORLD - F0 refinement, formants raise 30% | | 0.99 | 1.77 | 3.03 | 3.71 | 3.92 | 4.18 |

*Table A.8 Average P.563 results for recordings with "restaurant" noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 1.01 | 1.34 | 2.19 | 2.94 | 3.30 | 3.53 |
| Formants 20% | | 1.00 | 1.31 | 2.17 | 2.96 | 3.24 | 3.38 |
| Formants 30% | | 1.00 | 1.29 | 2.19 | 3.02 | 3.34 | 3.42 |
| Neutral | | 1.02 | 1.34 | 2.14 | 2.95 | 3.36 | 3.61 |
| PRAAT - F0 10% raise | | 1.01 | 1.36 | 2.33 | 3.10 | 3.53 | 3.95 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.01 | 1.38 | 2.42 | 3.15 | 3.58 | 3.81 |
| PRAAT - F0 10% raise, formants raise 20% | | 1.01 | 1.36 | 2.30 | 3.14 | 3.56 | 3.77 |
| PRAAT - F0 10% raise, formants raise 30% | | 1.01 | 1.29 | 2.30 | 3.11 | 3.53 | 3.71 |
| WORLD - F0 10% raise | | 1.02 | 1.56 | 2.71 | 3.76 | 4.20 | 4.43 |
| WORLD - F0 10% raise, formants raise 10% | | 1.03 | 1.61 | 2.78 | 3.77 | 4.11 | 4.29 |
| WORLD - F0 10% raise, formants raise 20% | | 1.03 | 1.65 | 2.96 | 3.92 | 4.19 | 4.29 |

| | | | | | | |
|---|---|---|---|---|---|---|
| WORLD - F0 10% raise, formants raise 30% | 1.02 | 1.60 | 2.81 | 3.78 | 4.14 | 4.24 |
| WORLD - F0 20% raise | 1.04 | 1.75 | 2.93 | 4.01 | 4.40 | 4.48 |
| WORLD - F0 20% raise, formants 3raise 0% | 1.04 | 1.76 | 3.08 | 4.06 | 4.44 | 4.49 |
| WORLD - F0 20% raise, formants raise 10% | **1.07** | 1.89 | 3.24 | 4.23 | **4.57** | 4.54 |
| WORLD - F0 20% raise, formants raise 20% | 1.05 | 1.76 | 3.15 | 4.18 | 4.35 | 4.46 |
| WORLD - F0 30% raise | 1.05 | 1.77 | 3.02 | 4.09 | 4.42 | 4.53 |
| WORLD - F0 30% raise, formants raise 10% | **1.07** | **1.95** | 3.27 | 4.27 | 4.50 | **4.59** |
| WORLD - F0 30% raise, formants raise 20% | 1.06 | 1.88 | **3.30** | **4.34** | 4.55 | 4.57 |
| WORLD - F0 30% raise, formants raise 30% | 1.04 | 1.78 | 3.16 | 4.21 | 4.49 | 4.55 |
| WORLD - F0 refinement | 1.04 | 1.54 | 2.63 | 3.63 | 4.07 | 4.37 |
| WORLD - F0 refinement, formants raise 10% | 1.03 | 1.51 | 2.62 | 3.61 | 3.96 | 4.17 |
| WORLD - F0 refinement, formants raise 20% | 1.01 | 1.54 | 2.71 | 3.60 | 3.97 | 4.09 |
| WORLD - F0 refinement, formants raise 30% | 1.02 | 1.50 | 2.69 | 3.72 | 4.11 | 4.17 |

*Table A.9 Average P.563 results for recordings with "street" noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 1.15 | 2.01 | 2.85 | 3.29 | 3.31 | 3.64 |
| Formants 20% | | 1.18 | 2.11 | 2.88 | 3.36 | 3.34 | 3.53 |
| Formants 30% | | 1.22 | 2.22 | 2.86 | 3.35 | 3.35 | 3.46 |
| Neutral | | 1.18 | 1.89 | 2.78 | 3.32 | 3.51 | 3.70 |
| PRAAT - F0 10% raise | | 1.18 | 2.03 | 2.99 | 3.52 | 3.65 | 4.00 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.20 | 2.15 | 3.11 | 3.59 | 3.68 | 3.96 |
| PRAAT - F0 10% raise, formants raise 20% | | 1.20 | 2.17 | 3.13 | 3.53 | 3.59 | 3.88 |
| PRAAT - F0 10% raise, formants raise 30% | | 1.20 | 2.09 | 2.99 | 3.43 | 3.55 | 3.78 |
| WORLD - F0 10% raise | | 1.22 | 2.31 | 3.56 | 4.11 | 4.29 | 4.55 |
| WORLD - F0 10% raise, formants raise 10% | | 1.31 | 2.47 | 3.65 | 4.18 | 4.24 | 4.37 |
| WORLD - F0 10% raise, formants raise 20% | | 1.36 | 2.49 | 3.59 | 4.22 | 4.25 | 4.38 |
| WORLD - F0 10% raise, formants raise 30% | | 1.40 | 2.45 | 3.52 | 4.12 | 4.13 | 4.28 |
| WORLD - F0 20% raise | | 1.32 | 2.44 | 3.79 | 4.43 | 4.47 | 4.53 |
| WORLD - F0 20% raise, formants 3raise 0% | | 1.38 | 2.72 | 3.87 | 4.46 | 4.51 | 4.51 |
| WORLD - F0 20% raise, formants raise 10% | | 1.39 | 2.70 | 4.06 | 4.61 | 4.55 | 4.59 |
| WORLD - F0 20% raise, formants raise 20% | | 1.41 | 2.68 | 3.95 | 4.56 | 4.50 | 4.54 |
| WORLD - F0 30% raise | | 1.30 | 2.60 | 3.79 | 4.49 | 4.62 | 4.57 |
| WORLD - F0 30% raise, formants raise 10% | | 1.35 | 2.86 | 4.06 | 4.63 | 4.55 | **4.65** |
| WORLD - F0 30% raise, formants raise 20% | | **1.44** | **2.91** | **4.11** | **4.65** | 4.55 | 4.56 |
| WORLD - F0 30% raise, formants raise 30% | | 1.42 | 2.85 | 3.97 | 4.59 | **4.57** | 4.59 |
| WORLD - F0 refinement | | 1.17 | 2.11 | 3.32 | 3.96 | 4.07 | 4.47 |
| WORLD - F0 refinement, formants raise 10% | | 1.16 | 2.17 | 3.36 | 3.94 | 3.94 | 4.23 |
| WORLD - F0 refinement, formants raise 20% | | 1.18 | 2.31 | 3.44 | 3.96 | 4.01 | 4.21 |
| WORLD - F0 refinement, formants raise 30% | | 1.20 | 2.27 | 3.38 | 4.03 | 4.11 | 4.21 |

*Table A.10 Average P.563 results for recordings with "exhibition" noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 0.98 | 0.99 | 1.65 | 2.70 | 2.95 | 3.42 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Formants 20% | 0.98 | 0.99 | 1.67 | 2.68 | 2.95 | 3.28 |
| Formants 30% | 0.98 | 1.00 | 1.68 | 2.72 | 3.06 | 3.26 |
| Neutral | 0.98 | 0.99 | 1.72 | 2.72 | 2.87 | 3.02 |
| PRAAT - F0 10% raise | 0.98 | 1.01 | 1.86 | 2.96 | 3.23 | 3.41 |
| PRAAT - F0 10% raise, formants raise 10% | 0.98 | 1.00 | 1.83 | 2.97 | 3.21 | 3.66 |
| PRAAT - F0 10% raise, formants raise 20% | 0.98 | 1.00 | 1.77 | 2.93 | 3.28 | 3.69 |
| PRAAT - F0 10% raise, formants raise 30% | 0.98 | 1.00 | 1.69 | 2.96 | 3.22 | 3.61 |
| WORLD - F0 10% raise | 0.98 | 1.04 | 2.18 | 3.37 | 3.43 | 3.43 |
| WORLD - F0 10% raise, formants raise 10% | 0.98 | 1.02 | 2.11 | 3.40 | 3.66 | 3.99 |
| WORLD - F0 10% raise, formants raise 20% | 0.98 | 1.04 | 2.14 | 3.47 | 3.78 | 4.23 |
| WORLD - F0 10% raise, formants raise 30% | **1.00** | 1.02 | 2.10 | 3.42 | 3.69 | 4.11 |
| WORLD - F0 20% raise | **1.00** | 1.05 | 2.34 | 3.70 | 3.76 | 3.73 |
| WORLD - F0 20% raise, formants 3raise 0% | 0.98 | 1.02 | 2.33 | 3.73 | 3.90 | 4.41 |
| WORLD - F0 20% raise, formants raise 10% | **1.00** | 1.05 | 2.51 | 3.91 | **4.02** | 4.26 |
| WORLD - F0 20% raise, formants raise 20% | **1.00** | 1.02 | 2.38 | 3.84 | 3.80 | 4.33 |
| WORLD - F0 30% raise | **1.00** | 1.05 | 2.36 | 3.66 | 3.60 | 3.77 |
| WORLD - F0 30% raise, formants raise 10% | **1.00** | 1.05 | **2.54** | **3.92** | 3.84 | 4.25 |
| WORLD - F0 30% raise, formants raise 20% | **1.00** | **1.06** | 2.51 | 3.86 | 3.87 | 4.39 |
| WORLD - F0 30% raise, formants raise 30% | **1.00** | 1.05 | 2.43 | 3.81 | 3.90 | **4.47** |
| WORLD - F0 refinement | 0.98 | 1.01 | 2.05 | 3.24 | 3.28 | 3.27 |
| WORLD - F0 refinement, formants raise 10% | 0.98 | 1.00 | 1.91 | 3.16 | 3.46 | 3.92 |
| WORLD - F0 refinement, formants raise 20% | **1.00** | 1.03 | 1.97 | 3.18 | 3.59 | 3.92 |
| WORLD - F0 refinement, formants raise 30% | **1.00** | 1.00 | 1.95 | 3.20 | 3.64 | 4.08 |

*Table A.11 Average P.563 results for recordings with pink noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 2.35 | 3.06 | 3.47 | 3.40 | 3.68 | 3.80 |
| Formants 20% | | 2.36 | 2.95 | 3.29 | 3.33 | 3.51 | 3.69 |
| Formants 30% | | 2.28 | 2.87 | 3.36 | 3.25 | 3.41 | 3.67 |
| Neutral | | 2.45 | 3.07 | 3.54 | 3.51 | 3.83 | 3.99 |
| PRAAT - F0 10% raise | | 2.94 | 3.50 | 3.88 | 3.86 | 3.96 | 4.17 |
| PRAAT - F0 10% raise, formants raise 10% | | 2.77 | 3.40 | 3.87 | 3.79 | 3.94 | 4.06 |
| PRAAT - F0 10% raise, formants raise 20% | | 2.62 | 3.38 | 3.76 | 3.77 | 3.88 | 3.98 |
| PRAAT - F0 10% raise, formants raise 30% | | 2.54 | 3.37 | 3.68 | 3.68 | 3.81 | 3.80 |
| WORLD - F0 10% raise | | 2.69 | 3.41 | 3.96 | 4.04 | 4.43 | 4.59 |
| WORLD - F0 10% raise, formants raise 10% | | 2.70 | 3.42 | 4.09 | 4.01 | 4.30 | 4.48 |
| WORLD - F0 10% raise, formants raise 20% | | 2.70 | 3.51 | 4.05 | 4.12 | 4.30 | 4.42 |
| WORLD - F0 10% raise, formants raise 30% | | 2.59 | 3.37 | 3.84 | 3.99 | 4.24 | 4.41 |
| WORLD - F0 20% raise | | 2.99 | 3.75 | 4.30 | 4.14 | 4.40 | 4.57 |
| WORLD - F0 20% raise, formants 3raise 0% | | 2.87 | 3.76 | 4.38 | 4.28 | 4.47 | 4.59 |
| WORLD - F0 20% raise, formants raise 10% | | 3.12 | **3.98** | 4.50 | **4.40** | 4.49 | 4.63 |
| WORLD - F0 20% raise, formants raise 20% | | 3.06 | 3.89 | 4.42 | 4.32 | 4.45 | 4.55 |
| WORLD - F0 30% raise | | 3.02 | 3.78 | 4.31 | 4.16 | 4.43 | 4.62 |
| WORLD - F0 30% raise, formants raise 10% | | **3.04** | 3.96 | **4.56** | 4.36 | **4.56** | **4.67** |
| WORLD - F0 30% raise, formants raise 20% | | **3.04** | 3.96 | 4.55 | 4.37 | 4.49 | 4.58 |
| WORLD - F0 30% raise, formants raise 30% | | 2.94 | 3.92 | 4.50 | 4.38 | 4.47 | 4.56 |
| WORLD - F0 refinement | | 2.55 | 3.21 | 3.82 | 3.94 | 4.32 | 4.51 |
| WORLD - F0 refinement, formants raise 10% | | 2.36 | 3.13 | 3.75 | 3.82 | 4.12 | 4.36 |

| | | | | | | |
|---|---|---|---|---|---|---|
| WORLD - F0 refinement, formants raise 20% | 2.31 | 3.16 | 3.72 | 3.83 | 4.11 | 4.32 |
| WORLD - F0 refinement, formants raise 30% | 2.36 | 3.17 | 3.75 | 3.81 | 4.13 | 4.39 |

*Table A.12 Average P.563 results for recordings with "car" noise*

| Type of modification | | P.563 | | | | | |
|---|---|---|---|---|---|---|---|
| | SNR: | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| Formants 10% | | 1.06 | 1.75 | 2.61 | 3.11 | 3.30 | 3.63 |
| Formants 20% | | 1.07 | 1.72 | 2.65 | 2.96 | 3.32 | 3.48 |
| Formants 30% | | 1.02 | 1.65 | 2.57 | 3.17 | 3.28 | 3.44 |
| Neutral | | 1.13 | 1.79 | 2.59 | 3.03 | 3.41 | 3.71 |
| PRAAT - F0 10% raise | | 1.15 | 2.08 | 2.92 | 3.37 | 3.66 | 3.90 |
| PRAAT - F0 10% raise, formants raise 10% | | 1.11 | 2.07 | 2.94 | 3.33 | 3.67 | 3.83 |
| PRAAT - F0 10% raise, formants raise 20% | | 1.04 | 1.93 | 2.90 | 3.47 | 3.66 | 3.76 |
| PRAAT - F0 10% raise, formants raise 30% | | 1.04 | 1.83 | 2.87 | 3.37 | 3.52 | 3.68 |
| WORLD - F0 10% raise | | 1.15 | 2.14 | 3.24 | 3.85 | 4.12 | 4.49 |
| WORLD - F0 10% raise, formants raise 10% | | 1.13 | 2.23 | 3.35 | 3.95 | 4.07 | 4.31 |
| WORLD - F0 10% raise, formants raise 20% | | 1.07 | 2.09 | 3.35 | 4.04 | 4.16 | 4.27 |
| WORLD - F0 10% raise, formants raise 30% | | 1.07 | 2.06 | 3.28 | 3.88 | 4.00 | 4.26 |
| WORLD - F0 20% raise | | **1.24** | 2.40 | 3.55 | 4.11 | 4.29 | 4.48 |
| WORLD - F0 20% raise, formants 3raise 0% | | 1.07 | 2.32 | 3.60 | 4.22 | 4.21 | 4.42 |
| WORLD - F0 20% raise, formants raise 10% | | 1.14 | **2.55** | 3.78 | 4.33 | **4.36** | 4.52 |
| WORLD - F0 20% raise, formants raise 20% | | 1.13 | 2.52 | 3.75 | 4.31 | 4.26 | 4.43 |
| WORLD - F0 30% raise | | 1.22 | 2.41 | 3.52 | 4.12 | 4.31 | 4.51 |
| WORLD - F0 30% raise, formants raise 10% | | 1.16 | 2.51 | 3.74 | 4.34 | 4.35 | **4.55** |
| WORLD - F0 30% raise, formants raise 20% | | 1.10 | 2.53 | **3.85** | 4.34 | 4.27 | 4.46 |
| WORLD - F0 30% raise, formants raise 30% | | 1.07 | 2.38 | 3.73 | **4.43** | 4.35 | 4.44 |
| WORLD - F0 refinement | | 1.11 | 2.02 | 3.01 | 3.67 | 4.04 | 4.38 |
| WORLD - F0 refinement, formants raise 10% | | 1.07 | 1.89 | 3.02 | 3.70 | 3.99 | 4.21 |
| WORLD - F0 refinement, formants raise 20% | | 1.02 | 2.01 | 3.08 | 3.71 | 3.97 | 4.21 |
| WORLD - F0 refinement, formants raise 30% | | 1.04 | 1.91 | 3.06 | 3.83 | 4.02 | 4.26 |

# Appendix B Detailed results of the main experiment

## Comparison of the speech improvements

The main experiment used the detailed calculations of P.563 value for the selected modification methods mentioned in Chapter 4.4. Every recording was modified and P.563 was calculated and – using its spectral features – the decision process was implemented using multi-layer perceptron. The averaged results for different types of noises, different values of SNR and different genders are presented in Table B.1. Best modifications are highlighted. The values presented are P.563 values calculated for the given types of recordings. The columns are, respectively:

- Neutral – neutral speech mixed with noise,
- AX - speech with F0 refinement using WORLD vocoder, with formants increased by 10%,
- CX - speech with F0 raised by 10% and refined, and with formants increased by 10%,
- Adaptive – results of the adaptive algorithm.

*Table B.1 Comparison of the results of different speech modification approaches*

| Noise type | SNR | Gender | Neutral | CX | AX | Adaptive |
|---|---|---|---|---|---|---|
| airport | -5 | f | 1.42 | **1.51** | 1.28 | 1.44 |
| airport | -5 | m | 1.06 | 1.09 | 1.01 | **1.14** |
| airport | 0 | f | 2.69 | **3.03** | 3.00 | 2.93 |
| airport | 0 | m | 1.39 | 1.99 | 1.63 | **2.57** |
| airport | 5 | f | 3.66 | 4.13 | **4.14** | 3.89 |
| airport | 5 | m | 1.95 | 3.05 | 2.72 | **3.77** |
| airport | 10 | f | 3.90 | 4.33 | **4.38** | 4.18 |
| airport | 10 | m | 2.32 | 3.63 | 3.23 | **4.05** |
| airport | 15 | f | 3.89 | **4.41** | 4.22 | 4.28 |
| airport | 15 | m | 3.03 | **4.09** | 3.80 | 3.94 |
| airport | 20 | f | 4.26 | 4.58 | 4.53 | **4.62** |
| airport | 20 | m | 3.40 | 4.28 | 4.03 | **4.29** |
| babble | -5 | f | 1.38 | **1.49** | 1.39 | 1.46 |
| babble | -5 | m | 1.10 | 1.33 | 1.18 | **1.40** |
| babble | 0 | f | 2.64 | **3.00** | 2.89 | 2.98 |
| babble | 0 | m | 1.63 | 2.37 | 1.95 | **2.84** |
| babble | 5 | f | 3.70 | 4.12 | **4.16** | 4.00 |
| babble | 5 | m | 2.18 | 3.35 | 2.93 | **3.96** |

| | | | | | | |
|---|---|---|---|---|---|---|
| babble | 10 | f | 4.15 | 4.63 | **4.68** | 4.39 |
| babble | 10 | m | 2.51 | 3.78 | 3.29 | **4.36** |
| babble | 15 | f | 4.05 | 4.33 | **4.47** | 4.37 |
| babble | 15 | m | 2.96 | 4.16 | 3.70 | **4.38** |
| babble | 20 | f | 4.24 | 4.51 | 4.49 | **4.58** |
| babble | 20 | m | 3.40 | 4.21 | 3.91 | **4.36** |
| car | -5 | f | **1.14** | 1.13 | 1.07 | 1.08 |
| car | -5 | m | 1.11 | 1.14 | 1.06 | **1.26** |
| car | 0 | f | 2.14 | **2.40** | 2.36 | 2.20 |
| car | 0 | m | 1.42 | 2.05 | 1.42 | **2.54** |
| car | 5 | f | 3.34 | 3.67 | **3.77** | 3.41 |
| car | 5 | m | 1.83 | 3.04 | 2.27 | **3.46** |
| car | 10 | f | 3.95 | 4.33 | **4.48** | 4.21 |
| car | 10 | m | 2.12 | 3.58 | 2.91 | **4.02** |
| car | 15 | f | 3.96 | 4.26 | **4.31** | 4.24 |
| car | 15 | m | 2.87 | 3.88 | 3.67 | **4.16** |
| car | 20 | f | 4.17 | 4.47 | 4.45 | **4.51** |
| car | 20 | m | 3.24 | 4.14 | 3.98 | **4.33** |
| exhibition | -5 | f | 1.00 | 1.00 | 1.00 | **1.00** |
| exhibition | -5 | m | 0.97 | 0.97 | 0.97 | **1.00** |
| exhibition | 0 | f | 1.02 | 1.05 | 1.04 | **1.08** |
| exhibition | 0 | m | 0.97 | 1.00 | 0.97 | **1.02** |
| exhibition | 5 | f | 2.23 | 2.48 | **2.50** | 2.39 |
| exhibition | 5 | m | 1.21 | 1.74 | 1.32 | **2.16** |
| exhibition | 10 | f | 3.51 | 3.84 | **3.95** | 3.82 |
| exhibition | 10 | m | 1.93 | 2.96 | 2.36 | **3.49** |
| exhibition | 15 | f | 3.49 | **4.09** | 4.08 | 3.92 |
| exhibition | 15 | m | 2.25 | 3.24 | 2.84 | **3.33** |
| exhibition | 20 | f | 3.35 | **4.39** | 4.35 | 4.00 |
| exhibition | 20 | m | 2.68 | **3.58** | 3.49 | 3.45 |
| pink | -5 | f | 3.21 | 3.11 | **3.14** | 3.07 |
| pink | -5 | m | 1.68 | 2.29 | 1.59 | **2.93** |
| pink | 0 | f | 3.93 | 3.96 | **4.05** | 3.87 |
| pink | 0 | m | 2.21 | 2.87 | 2.20 | **3.46** |
| pink | 5 | f | 4.47 | **4.60** | 4.57 | 4.51 |
| pink | 5 | m | 2.61 | 3.58 | 2.93 | **4.18** |
| pink | 10 | f | 4.02 | 4.28 | **4.30** | 4.26 |
| pink | 10 | m | 3.01 | 3.75 | 3.33 | **4.05** |
| pink | 15 | f | 4.22 | 4.51 | 4.50 | **4.57** |
| pink | 15 | m | 3.43 | 4.10 | 3.75 | **4.29** |

| | | | | | | |
|---|---|---|---|---|---|---|
| pink | 20 | f | 4.43 | 4.64 | 4.62 | **4.69** |
| pink | 20 | m | 3.56 | 4.32 | 4.11 | **4.49** |
| restaurant | -5 | f | 1.04 | 1.05 | **1.06** | 1.03 |
| restaurant | -5 | m | 1.00 | **1.01** | 1.00 | **1.01** |
| restaurant | 0 | f | 1.60 | **1.91** | 1.82 | 1.89 |
| restaurant | 0 | m | 1.09 | 1.30 | 1.20 | **1.70** |
| restaurant | 5 | f | 2.77 | 3.22 | **3.27** | 3.17 |
| restaurant | 5 | m | 1.50 | 2.35 | 1.98 | **2.81** |
| restaurant | 10 | f | 3.74 | 4.21 | **4.35** | 4.14 |
| restaurant | 10 | m | 2.17 | 3.34 | 2.87 | **3.87** |
| restaurant | 15 | f | 4.13 | 4.47 | **4.58** | 4.46 |
| restaurant | 15 | m | 2.58 | 3.74 | 3.35 | **4.35** |
| restaurant | 20 | f | 4.26 | 4.51 | 4.50 | **4.60** |
| restaurant | 20 | m | 2.97 | 4.07 | 3.85 | **4.39** |
| street | -5 | f | 1.27 | 1.46 | 1.29 | **1.50** |
| street | -5 | m | 1.08 | **1.15** | 1.02 | 1.11 |
| street | 0 | f | 2.40 | **2.96** | 2.75 | 2.92 |
| street | 0 | m | 1.38 | 1.97 | 1.60 | **2.18** |
| street | 5 | f | 3.52 | **4.10** | 4.08 | 4.06 |
| street | 5 | m | 2.05 | 3.20 | 2.64 | **3.62** |
| street | 10 | f | 4.17 | 4.63 | **4.67** | 4.55 |
| street | 10 | m | 2.47 | 3.73 | 3.20 | **4.36** |
| street | 15 | f | 4.24 | **4.54** | 4.41 | 4.47 |
| street | 15 | m | 2.78 | 3.93 | 3.48 | **4.47** |
| street | 20 | f | 4.26 | 4.53 | 4.53 | **4.59** |
| street | 20 | m | 3.14 | 4.20 | 3.94 | **4.43** |
| subway | -5 | f | 1.04 | 1.04 | **1.05** | 1.04 |
| subway | -5 | m | 1.09 | 1.02 | **1.03** | 1.01 |
| subway | 0 | f | 1.75 | 2.03 | **2.16** | 2.07 |
| subway | 0 | m | 1.10 | 1.84 | 1.38 | **2.12** |
| subway | 5 | f | 3.24 | 3.54 | **3.59** | 3.39 |
| subway | 5 | m | 1.72 | 2.79 | 2.28 | **3.47** |
| subway | 10 | f | 3.88 | 4.23 | **4.30** | 4.08 |
| subway | 10 | m | 2.14 | 3.58 | 3.00 | **3.99** |
| subway | 15 | f | 3.99 | 4.30 | **4.33** | 4.23 |
| subway | 15 | m | 2.70 | 3.90 | 3.47 | **4.38** |
| subway | 20 | f | 4.31 | 4.48 | 4.46 | **4.65** |
| subway | 20 | m | 3.05 | 4.11 | 3.98 | **4.33** |
| train | -5 | f | 2.27 | 2.41 | **2.43** | 2.39 |
| train | -5 | m | 1.22 | 1.68 | 1.22 | **2.13** |

| | | | | | | |
|---|---|---|---|---|---|---|
| train | 0 | f | 3.30 | 3.63 | **3.71** | 3.47 |
| train | 0 | m | 1.80 | 2.68 | 2.00 | **3.37** |
| train | 5 | f | 3.96 | 4.36 | **4.40** | 4.31 |
| train | 5 | m | 2.15 | 3.46 | 2.83 | **4.13** |
| train | 10 | f | 3.83 | **4.20** | **4.20** | 4.13 |
| train | 10 | m | 2.70 | 3.80 | 3.60 | **3.90** |
| train | 15 | f | 4,09 | 4.51 | 4.39 | **4.59** |
| train | 15 | m | 3,23 | 4.12 | 4.05 | **4.17** |
| train | 20 | f | 4,39 | 4.62 | 4.60 | **4.70** |
| train | 20 | m | 3,61 | 4.43 | 4.21 | **4.47** |

# Appendix C List of utterances used in experiments

The list of recorded sentences used in the experiments from the audio-visual corpus (Czyzewski *et al.*, 2017) for multimodal automatic speech recognition.

| **Utterance** |
|---|
| Wykonuj polecenia organów straży pożarnej i policji. |
| Kieruj się w stronę wyjścia ewakuacyjnego. |
| Proszę jak najszybciej opuścić budynek. |
| Zakaz korzystania z wind. |
| Proszę wezwać ochronę. |
| Czy wśród nas jest lekarz? |
| Gdzie znajduje się najbliższe wyjście ewakuacyjne? |
| Gdzie znajduje się sprzęt gaśniczy? |
| Czy ktoś potrafi udzielić pierwszej pomocy? |
| Czy została wezwana karetka pogotowia? |
| Nie ma zagrożenia, to nie jest pożar. |
| W prawym skrzydle budynku zostało wyłączone zasilanie. |
| Wszystkie pomieszczenia zostały przeszukane. |
| Winda uległa awarii, proszę poruszać się schodami. |
| Za chwilę nastąpi ewakuacja wszystkich osób z budynku. |

POLITECHNIKA GDAŃSKA

# Appendix D List of the author's publications


Korvel, G., Kąkol, K., Kurasova, O. and Kostek, B. (2020), "Evaluation of Lombard Speech Models in the Context of Speech in Noise Enhancement," IEEE Access, Vol. 8, pp. 155156–155170, doi: 10.1109/ACCESS.2020.3015421.

Kąkol, K., Korvel, G. and Kostek, B. (2020), "Improving Objective Speech Quality Indicators in Noise Conditions," Studies in Computational Intelligence, In Data Science: New Issues, Challenges and Applications. Studies in Computational Intelligence; Springer, Cham, Vol. 869, pp. 199–218, https://doi.org/10.1007/978-3-030-39250-5_9.

Kąkol K., Kostek B. (2016), "A study on signal processing methods applied to hearing aids"; „Przegląd metod przetwarzania dźwięku wykorzystywanych w aparatach słuchowych," Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej, vol. 51, pp. 71 - 76, 2016.

Korvel G., Kąkol K., Kostek B. (2021), "Similarity matrices applied to differentiate between Lombard and natural speech," The Journal of the Acoustical Society of America, Vol. 150 No. 4, pp. A114–A114, https://doi.org/10.1121/10.0007809.

Kostek, B. and Kąkol, K. (2018), "Improving the quality of speech in the conditions of noise and interference," The Journal of the Acoustical Society of America, Vol. 144 No. 3, pp. 1905–1905, https://doi.org/10.1121/1.5068349.

Kąkol K., Kostek B. (2018), "A Study on Improving Objective Quality Indicators of Speech Utterances in Noise Conditions"; „Poprawa obiektywnych wskaźników jakości mowy w warunkach hałasu," Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej, vol. 60, pp. 45 - 50, 2018, DOI: doi: 10.32016/1.60.09.

Korvel G., Kąkol K., Kostek B. (2019), Evaluation of Lombard speech models in the context of speech enhancement; 11th International Workshop on DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS, pp. 35 - 36, Druskininkai, Litwa, 28.11.2019 - 30.11.2019, DOI: 10.15388/DAMSS.11.2019.

Kąkol K. and Kostek B. (2016), "A study on signal processing methods applied to hearing aids," Signal Processing - Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, SPA, https://doi.org/10.1109/SPA.2016.7763616.

Kąkol K., Korvel G., Kostek B., (2018), "Analysis of Lombard speech using parameterization and the objective quality indicators in noise conditions"; 10th International Workshop "Data

analysis methods for software systems," DAMSS 2018, pp. 35 - 36, 29.11.2018 - 1.12.2018, DOI: Proc. DAMSS 2018: DOI: https://doi.org/10.15388/DAMSS.2018.1.

Kąkol, K., Korvel, G., Kostek, B. (2022), Noise profiling employing machine learning models, Journal Acoust. Soc. Amer. (*submitted to review*).

Korvel, G., Kąkol, K., Treigys P., Kostek B. (2022), Investigating the influence of noise interference on speech towards a future machine learning-based system for applying the Lombard effect automatically, ISMIS'2022, 26th International Symposium on Methodologies for Intelligent Systems Rende (Cosenza, IT), October 3-5, 2022, M. Ceci et al. (Eds.): ISMIS 2022, LNAI 13515, pp. 1–9, 2022, https://doi.org/10.1007/978-3-031-16564-1_38.

Korvel, G., Treigys P., Kąkol, K., Kostek B. (2022), "Does the Lombard effect cause problems for speech recognition, or is it a remedy for speech communication in noisy environments?," International Journal of Applied Mathematics and Computer Science (*submitted to review*).

Kąkol, K., Korvel, G., Tamulevičius, G., Kostek, B. (2022), "Detecting Lombard speech for speech-in-noise systems," Engineering Applications of Artificial Intelligence (*submitted to review*).