MULTIKONFERENCJA
ŚRODOWISKA
TELE- I RADIOKOMUNIKACYJNEGO

STRATEGIE TRENINGU NEURONOWEGO ESTYMATORA CZĘSTOTLIWOŚCI TONU KRTANIOWEGO Z UŻYCIEM GENERATORA SYNTETYCZNYCH SAMOGŁOSEK

TRAINING STRATEGIES OF NEURAL FUNDAMENTAL FREQUENCY ESTIMATOR USING A SYNTHETIC VOWELS GENERATOR

Marek Blok[1]; Jan Banaś[2]; Mariusz Pietrołaj[3]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, marek.blok@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jan.banas@pg.edu.pl
[3] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, mariusz.pietrolaj@pg.edu.pl

Streszczenie: **W wielu zastosowaniach telekomunikacyjnych pojawia się problem przetwarzania lub analizy sygnału mowy, w ramach którego, często w obszarze podstawowych algorytmów, stosuje się estymator częstotliwości tonu krtaniowego. Estymator rozpatrywany w tej pracy bazuje na neuronowym klasyfikatorze podejmującym decyzje na podstawie częstotliwości oraz mocy chwilowej wyznaczanych w podpasmach analizowanego sygnału mowy. W pracy rozważamy problematykę treningu tego estymatora, gdy trening odbywa się z użyciem sygnałów generowanych syntetycznie.**

Abstract: **In many telecommunication applications there is a need for a speech signal processing or analysis, within which the pitch tone frequency estimator is one of the common basic algorithms. The estimator considered in this paper is based on a neural classifier, whose decisions are driven by the instantaneous frequency and power determined in the sub-bands of the analyzed speech signal. In the paper, we consider the problems of selecting a training strategy for this estimator, when training is carried out with synthetically generated vowels.**

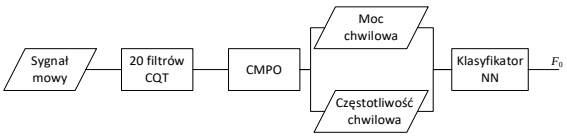
Słowa kluczowe: **estymacja częstotliwości tonu krtaniowego, klasyfikator neuronowy, częstotliwość chwilowa, trening sieci neuronowej.**

Keywords: **pitch estimation, neural classifier, instantaneous frequency, neural network training.**

1. WSTĘP

Jednym z dobrze rozpoznanych, lecz wciąż nierozwiązanych problemów w obszarze przetwarzania sygnałów mowy jest estymacja częstotliwości podstawowej tonu krtaniowego. Niezawodny algorytm realizujący to zadanie miałby szansę wspomóc rozwiązanie wielu problemów z zakresu przetwarzania sygnałów dźwiękowych, takich jak redukcja wymaganej przepustowości kanału transmisji sygnału poprzez kodowanie i resyntezę; poprawa jakości życia osób niesłyszących poprzez korektę intonacji, zniekształcanej przez aparat słuchowy bądź implant ślimakowy; separacja mówców w sygnale mowy dla zastosowań multimedialnych, takich jak telekonferencje. Niniejszy artykuł stanowi opis usprawnionego algorytmu IFE [1] - nowatorskiego podejścia do rozwiązania problemu estymacji częstotliwości podstawowej tonu krtaniowego. Opiera się ono o dwustopniowe przetwarzanie sygnałów dźwiękowych, którego schemat przedstawiono na rys. 1.

Pierwszy stopień stanowi przetwarzanie wstępne (*front-end*), realizujące ekstrakcję cech sygnału z rozdzielczością czasową rzędu pojedynczej próbki; następnie wyodrębnione cechy, zgrupowane w segmenty, stanowią wejście klasyfikatora neuronowego (*back-end*), realizującego przypisanie każdej próbki do jednej z predefiniowanych klas, odpowiadających spodziewanym wartościom częstotliwości tonu krtaniowego sygnału F_0. Klasyfikator został zaimplementowany jako sztuczna sieć neuronowa. Modyfikacją względem rozwiązania zaproponowanego w pracy [1] jest dodanie mechanizmów, które w ramach treningu pozwalają na generację syntetycznych danych treningowych na żądanie. Takie podejście umożliwia teoretycznie nieskończenie długi trening, co jest istotną zmianą, zwłaszcza w stosunku do treningów korzystających z rzeczywistych nagrań mowy, dla których istotnym problemem jest pozyskanie adnotacji dotyczącymi częstotliwości podstawowej tonu krtaniowego. Co więcej, ten sposób daje pewność co do referencyjnej wartości częstotliwości podstawowej tonu krtaniowego - w przeciwieństwie do istniejących korpusów z nagraniami, których adnotacje, powstałe na podstawie sygnału elektroglotograficznego, często obarczone są błędami wynikającymi również z samego mechanizmu powstawania dźwięcznych głosek [7].



Rys. 1. Schemat przetwarzania IFE

2. GENEROWANIE DANYCH

2.1. Synteza samogłosek

Kluczowym usprawnieniem IFE przedstawionym w niniejszym artykule jest metoda dostarczania danych treningowych do sieci neuronowej. Dzięki zastosowaniu syntezy samogłosek, dane mogą być generowane w czasie treningu, co pozwala na dostarczenie tak dużej ilości danych, jak jest żądane oraz eliminuje konieczność korzystania z dużych zasobów pamięci w czasie treningu.

$$A[k] = \left\{ \sin\left[\left(2\pi \frac{k \cdot F_0}{F_s} \right)^v \right] \right\}^{s^4} \quad (1)$$

Synteza samogłosek jest dwuetapowym procesem, który rozpoczyna się od poliharmonicznego generatora amplitudy, opisanego powyższym wzorem (1), w którym k opisuje indeks harmonicznej $F_k = k \cdot F_0$, ograniczonej w zakresie od losowo wygenerowanej F_0 do częstotliwości Nyquista $F_N = \frac{1}{2}F_s$, zależnej od szybkości próbkowania F_s. Dodatkowo wprowadzono losowe czynniki kształtujące rezonans samogłoski, v oraz s, których zakresy losowania zdefiniowano jako $v \in [0.15, 0.4]$ i $s \in [1.4, 2.5]$. Ostatecznie przebieg czasowy syntetycznej samogłoski jest generowany zgodnie ze wzorem

$$y[n] = \sum_{k=1}^{K} A[k] e^{j(k\hat{\varphi}[n] + \varphi_0[k])}, \quad (2)$$

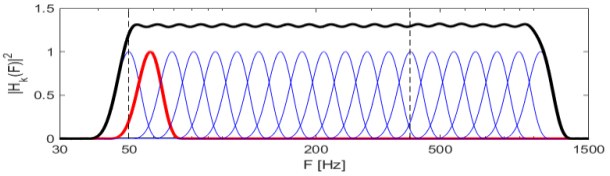
gdzie

$$\hat{\varphi}[n] = \frac{2\pi}{F_s} \sum_{m=0}^{n} F[m], \quad (3)$$

początkowa faza $\varphi_0[\cdot]$ jest losowana a chwilowa częstotliwość podstawowa $F[\cdot]$ ma losową wartość początkową i nieliniowo zmienia się w czasie. Kolejny etap addytywnej syntezy to dodanie zniekształceń typu shimmer i jitter [6] oraz dodanie szumu białego o kontrolowanym poziomie, reprezentującego naturalne zakłócenia powstające w ramach wypowiedzi jak i przy rejestracji sygnału, pozwalającym na regulację wynikowego stosunku poziomu sygnału użytecznego do szumu (SNR).

2.2. Przetwarzanie wstępne - ekstrakcja cech

Zarówno syntetyczne samogłoski wygenerowane w sposób opisany powyżej, jak i nagrania rzeczywistej mowy, zostają podane wstępnemu przetwarzaniu, którego celem jest ekstrakcja cech sygnału na potrzeby sieci neuronowej. Celem jest uzyskanie wartości chwilowej mocy i częstotliwości dla każdej kolejnej próbki analizowanego sygnału, w każdym z 20 podpasm częstotliwościowych. Przetwarzanie to obejmuje dwa etapy – filtrację oraz obliczanie zespolonej mocy wzajemnej.



Rys. 2. Bank filtrów przetwarzania wstępnego

Pierwszy etap polega na podaniu sygnału wejściowego na 20 filtrów o stałej dobroci, częstotliwościach środkowych rozłożonych w zakresie od 50Hz do 1050Hz. Są to filtry Hilberta o skończonej odpowiedzi impulsowej, generujące na wyjściu sygnały o wartościach zespolonych. Filtry te zostały zaprojektowane metodą jądra CQT (*Constant-Q Transform*) [5] z użyciem okna Blackmana, jak pokazano na rys. 2.

W każdym z podpasm wydzielonych w wyniku powyższej filtracji, stosowany jest operator zespolonej mocy wzajemnej (*complex mutual power operator - CMPO*) zdefiniowany jako:

$$y_{\mathrm{CMPO}}[n] = x[n] \cdot x^*[n-1], \quad (4)$$

gdzie $x[n]$ próbką wyjściową z jednego z filtrów, a $x^*[n]$ jej wartością sprzężoną. Na podstawie CMPO można łatwo obliczyć moc chwilową

$$P_i[n] = |y_{\mathrm{CMPO}}[n]| \quad (5)$$

i częstotliwość chwilową

$$F_i[n] = \arg(y_{\mathrm{CMPO}}[n]). \quad (6)$$

Tak obliczone wartości nie oferują wystarczającej dokładności by stanowić samodzielny estymator częstotliwości podstawowej tonu krtaniowego, niosą one jednak wystarczająco dużo informacji dotyczącej dominujących zakresów częstotliwości (moce chwilowe) i prawdopodobnych chwilowych wartości częstotliwości (częstotliwości chwilowe), żeby użyć ich do dalszego przetwarzania z użyciem sztucznej sieci neuronowej.

3. SIEĆ NEURONOWA

Przedstawiona metoda określania podstawowej częstotliwości tonu krtaniowego bazuje na algorytmie sztucznej inteligencji zaimplementowanym w postaci sieci neuronowej. Trenowany model wykorzystuje architekturę wielowarstwowego perceptronu, zawierającego dwie ukryte warstwy liczące odpowiednio 250 i 500 neuronów [1]. Topologia warstwy wejściowej jest podyktowana rozmiarem danych zwracanych przez generator danych treningowych, zakłada on 40 wartości odpowiadającym częstotliwości i energii każdej próbki sygnału.

3.1. Trening i klasyfikacja

Inferencja sieci neuronowej przyporządkowuje każdą z analizowanych próbek sygnału jako jedną z 351 predefiniowanych wartości częstotliwości. Odpowiadają one przedziałom o logarytmicznej dystrybucji w zakresie od 50 do 400 Hz. Decyzja o zastosowaniu skali logarytmicznej została podyktowana charakterystyką sygnału mowy i większym skupieniu F_0 w niższych przedziałach częstotliwości [1]. Ważnym wyróżnikiem prezentowanej metody jest użycie wyłącznie syntetycznie generowanych danych do treningu i testowania sieci neuronowej. Tylko ostateczna ewaluacja i porównanie zastosowanego rozwiązania bazuje na rzeczywistych sygnałach mowy takich jak baza danych Keele Pitch [2]. Zastosowany mechanizm generacji danych pozwala na wykorzystanie nowego zestawu treningowego na każdą z epok treningu.

W przypadku prezentowanych wyników poziom SNR generowanych sygnałów był analogiczny do danych wejściowych zastosowanych w IFE i wynosił 40 dB [1]. W zależności od wariantu eksperymentu, unikalny zestaw treningowy pojedynczej epoki zawierał od 5000 do 50000 próbek, co w najdłuższym wariancie odpowiada około 7 godzinom ciągłego sygnału audio. Przed podaniem parametrów na warstwę wejściową sieci były one poddawane procesowi normalizacji w zakresie od 0 do 1 w dziedzinie częstotliwości i energii. Syntetyczny zestaw walidacyjny obejmował 1000 syntetycznych głosek i był używany wyłącznie do zweryfikowania dokładności trenowanej sieci.

3.2. Parametryzacja sieci

Proces treningu obejmował zróżnicowaną liczbę epok z uwzględnieniem podziału danych wejściowych na zestawy po 128 próbki (*batch size*). W celu aktywacji wykorzystano funkcję ReLU. Obliczenie funkcji straty opierało się na entropii skrośnej (*cross entropy*) i klasyfikacji przy pomocy funkcji softmax. Zastosowany parametr szybkości uczenia (*learning rate*) wynosił 10^{-4}. Omawiane rozwiązanie z uwzględnieniem treningu, walidacji i ewaluacji zostało zaimplementowane przy pomocy języka Python i środowiska PyTorch [3].

4. EKSPERYMENTY

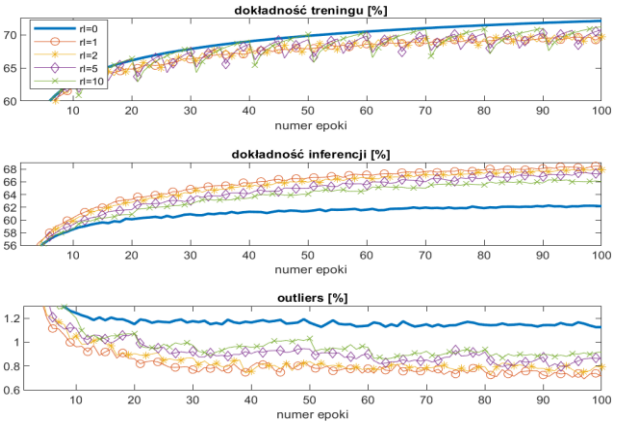
Wykorzystanie syntetycznych danych treningowych generowanych na żądanie pozwala na użycie w każdej kolejnej epoce nowego ich zestawu. Pozwala to uniknąć przeuczenia sieci dla danego zestawu treningowego. Problem ten jest szczególnie istotny, gdy dostępne dane treningowe są ograniczone. Przykładowo użyta do walidacji metody IFE baza Keele Pitch zawiera 10 wypowiedzi, 5 żeńskich i 5 męskich, o łącznym czasie trwania 337 sekund, z czego 159 sekund to segmenty dźwięczne opatrzone informacją o częstotliwości tonu krtaniowego. Dla porównania do treningu sieci w [1] użyto 30 000 półsekundowych segmentów (syntetycznych głosek).

W ramach prac nad metodą IFE zauważyliśmy, że chociaż wielokrotne podawanie tego samego zestawu danych prowadzi do przetrenowania sieci, to jednak, jeżeli liczba segmentów jest dostatecznie duża, kilkukrotne wykorzystanie tego samego zestawu danych pozwalało na uzyskanie poprawy jakości inferencji. Maksymalny rozmiar zestawu treningowego ogranicza jednak rozmiar dostępnej pamięci, zwłaszcza, że na wyjściu przetwarzania wstępnego na każdą próbkę sygnału treningowego otrzymujemy 40 wartości analizowanych przez sieć.

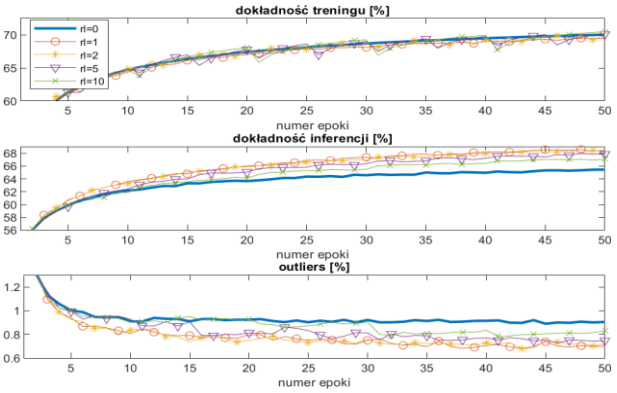
Problem ten skłonił nas do przeanalizowania efektywności rozwiązania, w ramach którego syntetyczne dane treningowe są generowane w mniejszych segmentach na potrzeby treningu kolejnych epok. W trakcie eksperymentów zmienialiśmy zarówno liczbę syntetycznych głosek używanych do treningu w ramach danej epoki N_{seg} (5000, 10000, 25000 oraz 50000), jak i częstość ponownego użycia danych treningowych w kolejnych epokach rl, gdzie 0 oznacza użycie tych samych danych we wszystkich epokach, a większa wartość określa, co ile epok generowany jest nowy sygnał. Do walidacji wytrenowanej sieci użyto danych syntetycznych złożonych z 1000 półsekundowych głosek o SNR równym 40dB.

Jak można zauważyć na rys. 3-6 dla małego N_{seg} (rys. 3 i 4) zmiana zestawu danych treningowych skutkuje skokowym pogorszeniem dokładności obserwowanej podczas treningu oraz drobnemu skokowi w górę w dokładności inferencji. Ponadto wielokrotne użycie tego samego zestawu danych treningowych skutkuje zwiększeniem się udziału w estymatach F_0 wartości odchylonych od wartości referencyjnej o więcej niż 20% (*outliers*), który przy przełączaniu zestawu skokowo spada.

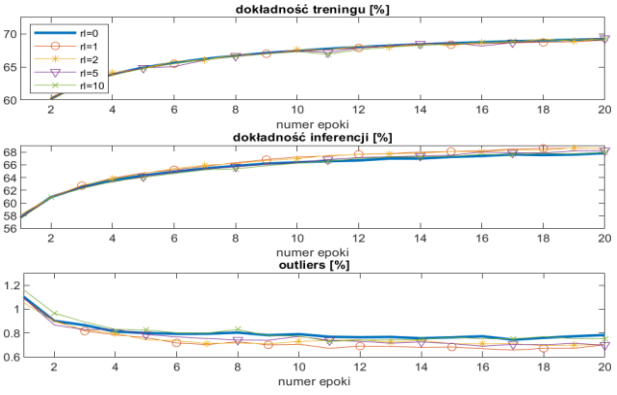
Biorąc pod uwagę jakość inferencji wytrenowanej sieci, to najlepsze efekty, dla krótkich zestawów danych treningowych, daje ciągła wymiana danych treningowych. Różnice pomiędzy strategiami treningu zacierają się wraz ze wzrostem wielkości zestawu treningowego (rys. 5 i 6). Dla N_{seg} = 50 000 co prawda ponownie najlepsza strategia to generowanie nowych danych co epoka (rl = 1), ale różnice nie są tym razem tak duże w porównaniu do innych strategii. Ta wielkość zestawu jest bliska maksymalnej wielkości segmentu pozwalającej na trening na komputerze z 64GB RAMu, a jednocześnie zbliżone wyniki uzyskano dla innych wartości N_{seg}, w tym 5000, dla rl = 1, jednak z nieco zwiększoną liczbą outlierów.



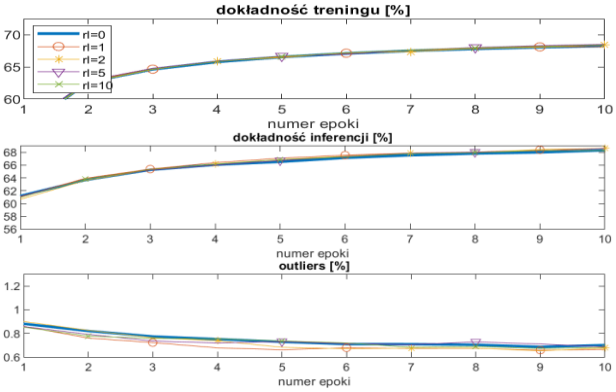
Rys. 3. Dokładność treningu i inferencji oraz procent znacznych błędów dla treningu z N_{seg}= 5 000



Rys. 4. Dokładność treningu i inferencji oraz procent znacznych błędów dla treningu z N_{seg}= 10 000



Rys. 5. Dokładność treningu i inferencji oraz procent znacznych błędów dla treningu z N_{seg}= 25 000



Rys. 6. Dokładność treningu i inferencji oraz procent znacznych błędów dla treningu z N_{seg}= 50 000

Lekka przewaga treningu z największym N_{seg} prawdopodobnie wynika z tego, że tasowanie danych treningowych silniej zaburza korelację pomiędzy danymi dla kolejnych próbek sygnału. Sygnały te co prawda cechują się zmiennymi parametrami, ale zachowują ciągłość tych zmian w ramach poszczególnych głosek, na które składa się aż 4000 próbek.

Trening z N_{seg} = 5 000 przy ciągłej resyntezie danych pozwala na trening ze znacznie mniejszymi wymaganiami pamięciowymi z kolei trening z $N_{seg} \geq$ 50 000 pozwala na wielokrotne wykorzystanie raz wygenerowanych danych treningowych.

W tab. 1 i 2 przedstawiono wyniki pomiaru dokładności estymacji F_0 dla wypowiedzi z bazy Keele Pitch. Jest to procent estymat mieszczących się w 5% przedziale wokół wartości referencyjnej. W zestawieniu uwzględniono cztery strategie treningu: (a) N_{seg} = 5 000 oraz rl = 0, (b) N_{seg} = 5 000 oraz rl = 1, (c) N_{seg} = 50 000 oraz rl = 0 i (d) N_{seg} = 50 000 oraz rl = 1.

Tab. 1. Procent poprawnych estymat F_0 dla tolerancji ±5% dla wypowiedzi żeńskich

	f1	*f2*	*f3*	*f4*	*f5*
(a)	91,1	95,4	93,2	91,3	94,5
(b)	91,8	95,6	93,0	92,4	96,4
(c)	91,2	95,6	93,5	92,3	94,5
(d)	91,1	95,5	92,7	92,1	94,0

Tab. 2. Procent poprawnych estymat F_0 dla tolerancji ±5% dla wypowiedzi męskich

	m1	*m2*	*m3*	*m4*	*m5*
(a)	69,2	85,9	91,6	87,5	85,0
(b)	71,6	87,1	93,5	88,4	86,9
(c)	71,5	86,2	92,6	89,3	87,9
(d)	71,8	86,1	92,5	89,8	86,8

W tym przypadku można zauważyć, że trening dla krótkiego segmentu z ciągłą generacją nowego zestawu (b) poza sygnałem f3, dla którego odnotowano niewielkie pogorszenie, skutkuje poprawą w porównaniu do ciągłego wykorzystania tego samego zestawu danych treningowych (a). Średnio wyniki są lepsze o 0,7% dla głosów żeńskich oraz 1,7% dla głosów męskich, z najlepszą poprawą dla głosu m1 (najtrudniejszego przypadku) wynoszącą 2,4%. Z kolei dla długich segmentów (c i d) różnice są wyraźnie mniejsze, a wyniki są zbliżone (wahania w zakresie 1%) do wyników z wariantu (b), z wyróżniającymi się wynikami dla wypowiedzi f5, dla której strategia (b) dała prawie 2% poprawę.

5. PODSUMOWANIE

W ramach przedstawionych w pracy badań zweryfikowano różne strategie treningu sztucznej sieci neuronowej estymującej częstotliwość tonu krtaniowego z wykorzystaniem możliwości generowania danych treningowych na żądanie. Warto zwrócić uwagę na to, że wykorzystanie w treningu syntetycznych głosek, z jednej strony jest skuteczne, co potwierdzają przedstawione wyniki, a z drugiej strony jest konieczne, bo trening sieci wymaga kilku godzin dźwięcznych segmentów mowy opatrzonych wiarygodną wartością F_0, co stanowi dużym problem. Jednocześnie nawet korzystając z dużego zestawu danych treningowych problem stanowią wymagania pamięciowe, które można ograniczyć realizując trening dla mniejszych zestawów danych generowanych na bieżąco w trakcie treningu. Jak pokazano, mały zestaw danych treningowych może być skutecznie użyty do treningu, o ile wymiana danych treningowych następuje dla każdej epoki. Taka strategia pogarsza osiągi obserwowane bezpośrednio w ramach treningu, ale poprawia wyniki inferencji, które są porównywalne z wynikami uzyskiwanymi dla treningu z wykorzystaniem dużych segmentów danych.

LITERATURA

[1] Blok, Marek, Jan Banas, and Mariusz Pietrolaj. 2021.“IFE: NN-Aided Instantaneous Pitch Estimation.”. 14th International Conference on Human System Interaction (HSI).

[2] “Keele Pitch Database”. https://lost-contact.mit.edu/afs/nada.kth.se/dept/tmh/corpora/KeelePitchDB/. Accessed April 23, 2022.

[3] “Pytorch.” PyTorch. https://pytorch.org/. Accessed April 23, 2022. – C

[4] Rosenblatt, F. 1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. Psychological Review 65, no. 6: 386–408.

[5] Schörkhuber, Christian, and Anssi Klapuri. 2010. “Constant-Q Transform Toolbox for Music Processing”. 7th Sound and Music Computing Conference.

[6] Teixeira, João Paulo, Carla Oliveira, and Carla Lopes. 2013. “Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters”. Procedia Technology 9: 1112–22.

[7] Veprek, Peter, and Michael S. Scordilis. 2002. "Analysis, enhancement and evaluation of five pitch determination techniques". Speech Communication 37.3-4: 249-270.