2014 ACTA UNIVERSITATIS CAROLINAE PAG. 69–78
PHILOLOGICA 1 / PHONETICA PRAGENSIA XIII

THE FOURIER AND WAVELET TRANSFORMS IN SPEECH PROCESSING: THE CASE OF HARMONICITY

JANA HERANOVÁ, JAKUB KAPRÁL,
VERONIKA POJAROVÁ

ABSTRACT

This paper compares two methods of signal analysis – the Fourier transform as the most commonly used method in speech processing and the wavelet transform as a rather new approach to signal analysis. The potential of the wavelet transform in speech processing has not been fully explored yet. We have attempted to confront both methods on the example of the harmonicity measure (HNR). The Praat environment, which utilizes the Fourier transform, was used to provide reference values of HNR. For the MATLAB environment we proposed a method of HNR estimation which makes use of the wavelet transform. It was expected that the proposed approach would yield similar results as the established HNR estimation in Praat. This hypothesis has not been confirmed, as shown by the results from the HNR measurement in MATLAB which do not reflect the aperiodicity of the speech signal caused by F0 perturbations.

Key words: HNR, harmonicity, Fourier transform, wavelet transform

1. Introduction

Human speech is a type of mechanical wave, which is transmitted through an elastic medium. It changes with time, hence in its raw form it can be regarded as a time-domain based signal. The simplest way to depict the speech signal is through an oscillogram (waveform) where the amplitude of a signal is displayed as a function of time (see Figure 1). Such a representation shows how the signal changes in time. For the purposes of an acoustic analysis this representation is, however, not sufficient as it does not provide us with the information about the frequency content of the signal. To acquire a different representation of a signal mathematical transformations are used. In this article we are going to discuss the Fourier and the wavelet transforms. The Fourier transform is the most commonly used transformation in speech processing, whereas the wavelet transform is quite popular in image processing and remains largely unexplored in speech research.

The basic principle that lies behind the Fourier and wavelet transforms was postulated in the 19th century and today is referred to as the Fourier Theorem. It states that a complex wave can be described as a sum of fundamental components. These components

together form a representation which provides a different view of the signal. In the Fourier transform the complex signal is decomposed into sine and cosine waves. The result of such analysis is a frequency spectrum where the amplitude is displayed as a function of frequency. In the spectrum we achieve a perfect frequency resolution without any information about the time dimension, whereas in the oscillogram the signal is displayed with perfect time resolution but no information about the frequency components. The speech signal changes in time continuously, and therefore a static frequency spectrum captured for one specific moment does not include sufficient information for signal analysis. An alternative visualization which reflects the changes of spectrum in time introduces the spectrogram (see Figure 1). It contains both information about frequency and time but the resolution in the frequency and the time domain is compromised. This is due to the uncertainty principle which states that there is a limit to the precision with which energy and time can be measured simultaneously (Zimmermann, 2002: 121–126).

Compared to the Fourier transform the wavelet transform is a rather new method of signal analysis. They both work with the same principle of signal decomposition but in the wavelet transform the signal is decomposed as modified versions of a mother wavelet.

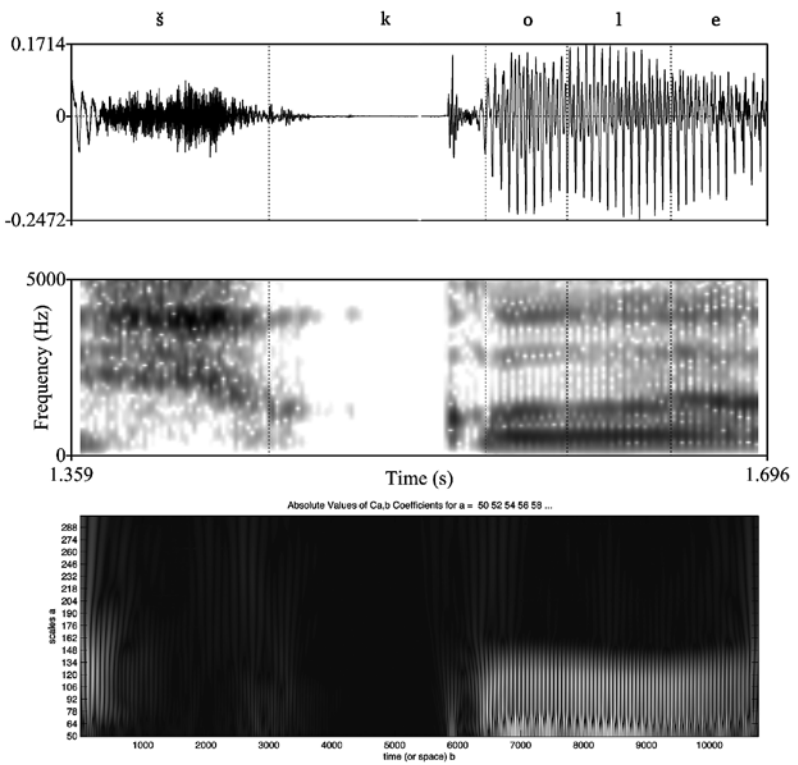

Figure 1. Visualisation of the word "škole". At the top oscillogram (waveform). In the middle wide-band spectrogram traditionally used in phonetic research. At the bottom scalogram – the lighter the color the higher the correlation between the wavelet and the signal.

Wavelets are functions with limited energy and limited time duration, and can therefore be regarded as short oscillations. There are many wavelet families (e.g. Haar, Daubechies, Morlet, Meyer or Mexican hat wavelet) containing mother wavelets with different shapes and properties. The selection of the wavelet depends on the purpose of the analysis and on the properties of the signal. Selesnick (2007) informs that there are two basic types of wavelet transform: one type is designed for reversible transformation, the other for signal analysis. Reversible transformation allows the original signal to be easily reconstructed. In image processing this feature is used for image compression and cleaning. The computed wavelet transform of an image is modified and subsequently the transformation is reversed and a new modified image is produced. The result of wavelet transform also depends on the properties of the signal. Zimmermann (2002: 139) reports that the shape of a wavelet should preferably resemble the input signal. In the wavelet transform the signal is analysed using a mother wavelet with different scales to estimate the correlation between the wavelet and the signal in time. The result of this analysis is a scalogram, a plot depicting the degree of correlation between the wavelet and the signal in time for various scales of the wavelet (see Figure 1). This visualization is similar to the spectrogram but instead of the changes of spectrum in time we follow the correlation between the wavelet and the signal. By choosing a different approach to the decomposition of the signal the wavelet transform avoids the problem of time-frequency resolution that we encounter in the Fourier transform. It therefore appears that signal analysis using the wavelet transform might be quite convenient for speech signal processing, especially for analysis of non-stationary phenomena. A more detailed account of both the Fourier and the wavelet transforms with its technical aspects can be found in Zimmermann (2002) or in Mallat (1998).

The Fourier analysis is widely used and it is by far the most popular transformation in speech processing regardless of its disadvantages which include the already mentioned time-frequency resolution, as well as the windowing effect and spectral leakage. The effect of using window functions to compute spectral estimates is described by Cox et al. (1989). The authors explain that for signal analysis we use a segment of data defined by a window function. The application of the window function results in the spreading of spectral energy called spectral leakage. The influence of spectral leakage can be minimized by usage of a tapered window function. However, it has impact on the spectral resolution (compare also Zimmermann, 2002: 71–83). Compared to the mentioned issues the wavelet transform looks promising with regard to speech analysis. In this article our aim is to explore the potential of the wavelet transform for phonetic research. We attempt to apply the Fourier and the wavelet transforms for speech analysis, specifically for the measuring of harmonicity.

Harmonicity (harmonics-to-noise ratio or HNR) is a measure expressing the overall acoustic periodicity of a voice signal (Murphy, 2006). It quantifies the ratio between the harmonic and the noise components in the speech signal in terms of dB. The noise components may be a result of the additional noise produced at the glottis during phonation (Awan and Frenkel, 1994) and/or of temporal and amplitude perturbations of the fundamental frequency (Murphy, 2006; Qi and Hillman, 1997). The algorithms for HNR estimation were proposed in the 1980's. HNR was originally developed as an objective measure to capture the perceptual properties of voice, its main purpose being to iden-

tify voice pathology. Later, other applications of HNR have also been explored: Ferrand (2002) suggests that HNR is a sensitive indicator of voice aging. Heranová and Skarnitzl (2011) examined HNR as an indicator of speech sound boundaries. The measuring of harmonicity has been repeatedly revised and new methods have been developed. In the current research we distinguish time- and frequency-domain based methods of HNR measurement. The time-domain based methods compute HNR from the acoustic waveform whereas the frequency-domain based methods use for HNR estimation spectral or cepstral representations of the signal (Qi and Hillman, 1997). The first time-domain based approach to HNR estimation was proposed by Yumoto et al. (1982). The first step in this approach is to gain one period of an average wave, which defines the energy of the harmonic component. Such a wave is calculated as the mean of more successive periods. The noise component is represented by the energy of variance between the average wave and the individual periods of the signal. The weakness of the time-domain based methods is the need to delimit the individual periods (Murphy, 2006). A solution to such a signal pre-processing is introduced in the form of the frequency-domain based methods, which compute HNR from the spectrum, calculating the ratio between the energy of harmonic and noise components in the signal. The energy of harmonic components is constituted by the energy at harmonic locations, and the energy of noise components consists of "between harmonic" estimates. The drawback of these methods is the problematic estimation of noise at harmonic locations (Murphy, 2006). A popular method of HNR measurement is introduced in the software for speech analysis – Praat (Boersma and Weenink, 2011). It is based on the autocorrelation of the signal, a method used for the detection of periodicity, in other words for F0 estimation. It calculates the correlation between a small chunk of the signal and the signal itself over a range of possible period durations (Johnson, 2003: 30). The period duration, which produces the maximum of the correlation function, represents the best F0 candidate. The relative height of this maximum represents the degree of periodicity of the signal and is used for the estimation of the relative power of the periodic component. Furthermore, because the autocorrelation of the signal equals the sum of the autocorrelations of its parts, the relative power of the noise component is calculated as a complement to the power of the periodic component (Boersma, 1993).

The aim of this paper is therefore to compare two methods for research of harmonicity which use a different approach to signal analysis. The first method is employed in Praat (Boersma and Weenink, 2011) and builds on the Fourier transform and the autocorrelation method to compute HNR. The second method for HNR estimation uses the wavelet transform in the MATLAB environment (Mathworks, 2012). The HNR estimation using wavelet transform was successfully implemented, e.g. in the research of Zhao et al. (2003) who investigated the properties of pathological voices. Our research concentrates on other than pathological purposes of HNR and is therefore carried out on the material aquired from healthy speakers. This suggests that we can expect very fine differences in the HNR values. This paper presents an attempt to propose an algorithm for HNR estimation which would yield similar results as the well-established method in Praat.

2. Method

For the experiment we used recordings of 10 native Czech speakers, 5 male and 5 female speakers aged 20–23, created in a sound-proof booth. The speech sound boundaries were set manually. As Ladefoged (1996: 22) reports, most of the frequencies that are of interest for speech analysis are below 8000 Hz. The most important acoustic cues for vowel analysis are the first five formants that are usually located below 5000 Hz. Therefore we can afford to resample the signal with a lower sampling frequency. As a result the informational content of the signal is decreased without omitting the components that are energetically significant for the analysis. On this account we resampled the material from 32 kHz to 9984 Hz. The wavelet decomposition, as implemented in the GUI resynt3 developed at the Institute of Phonetics, required the use of a sampling frequency equalling a multiple of 64. The sampling frequency 9984 Hz was chosen as the best candidate closing to the value of 10,000 Hz. For the analysis we used only vowels with duration greater than 80 ms, which were extracted from the original recordings using a rectangular window. The vowel duration criterion was motivated by the autocorrelation-based HNR measurement in Praat. Boersma (1993) informs that for HNR estimation the autocorrelation function requires a window at least 6 periods long. The standard setting of minimum F0 to 75 Hz then necessitates an analysis window with the length of 80 ms. In total, 791 samples of short and long vowels and diphthongs – specifically [ɪ i: e e: a a: o o: u u: a͡u o͡u] – were extracted and used for HNR measurement.

The aim of this experiment was to confront the Fourier and the wavelet analysis on the example of HNR. As reference values were used the results of the well-established AC-based method for HNR estimation in Praat. The harmonicity values were measured using cross-correlation method with standard settings. Every sound object was transformed into a harmonicity object, which represents the degree of acoustic periodicity. From the harmonicity object the mean values of HNR were retrieved.

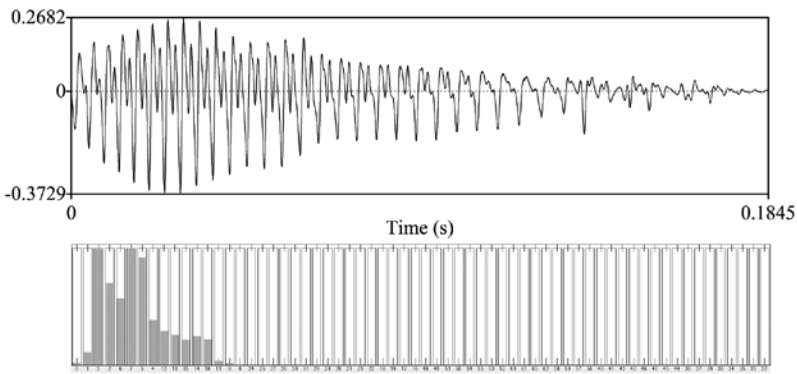

Figure 2. Example of waveform (top) and its energy distribution (bottom) for the vowel [u]. The distribution of energy was estimated in GUI resynt3 in the MATLAB environment. The frequencies from 0–4992 Hz are divided into 64 frequency bands with a width of 78 Hz. The distribution of energy within the bands is indicated by the height of the grey columns.

To get wavelet-based HNR estimates we employed the method of the wavelet packet decomposition that was implemented in the GUI resynt3 in the MATLAB environment. The wavelet packet decomposition is an extension of the discrete wavelet transform that allows a reversible transformation of the signal, i.e. signal decomposition and lossless reconstruction. We used decomposition of the 6th level where the signal is analysed into 2^6 evenly wide frequency bands. The decomposition of the material sampled at 9984 Hz provided us with 64 frequency bands with a width of 78 Hz (see Figure 2). The decomposition was done using discrete Meyer wavelet as the basis function.

Based on the energy distribution within the frequency bands we appointed 3 different frequency intervals (0–1248 Hz, 0–1326 Hz and 0–1404 Hz) where the most energy was distributed to represent the energy of the periodic component. The different frequency intervals were selected to check whether there is any significant dependency on a specific frequency range to represent the periodic energy of the signal. After the decomposition of the signal we reconstructed for each sound its periodic component using the GUI resynt3. The reconstruction was done for 3 sets of data based on 3 different frequency intervals representing the periodic component of the signal. The computation of HNR was based on the common assumption that the signal consists of the periodic and the noise component. Analogously the energy of the signal equals the sum of the energies of its parts. Hence, in our approach the energy of the noise component was calculated as a complement to the energy of the periodic component. The wavelet-based HNR estimate was calculated as a logarithmic ratio between the energy of the periodic and the noise component. Though we are aware of the fact that this may be a very crude measure of the signal's harmonic content, we will continue to use the term HNR even for the ratio calculated based on this wavelet decomposition.

3. Results and discussion

The vowels are speech sounds with periodic character. Their periodic component is dominant and in healthy speakers it prevails over noise. This also applies to the distribution of energy in vowels – the most energy is concentrated in harmonic locations. Based on these considerations we measured 3 sets of data with the wavelet-based approach in MATLAB. In each set the periodic component was represented by a different frequency band of the signal (0–1248 Hz, 0–1326 Hz and 0–1404 Hz) and the noise component was calculated as the complement. This selection was motivated by the visually observed distribution of energy in the vowel samples (see Figure 2).

First we tested the 3 sets of data against each other. A significant correlation ($p < 0.05$) between the 3 groups was discovered. This means that the setting of boundary, which in our approach separates the periodic and the noise part of the signal, in the interval between 1248–1404 Hz, does not significantly influence the value of HNR.

Afterwards we tested one set of wavelet-based HNR estimates, where the periodic component was represented by the frequency band 0–1248 Hz, against the values retrieved in Praat by the AC-based approach. Both data sets were analysed with ANOVA (see Figure 3). The relationship between the harmonicity and vowel's quality proved to be significant both for the AC-based ($F(11,779) = 15.56$; $p < 0{,}001$) and the wavelet-based

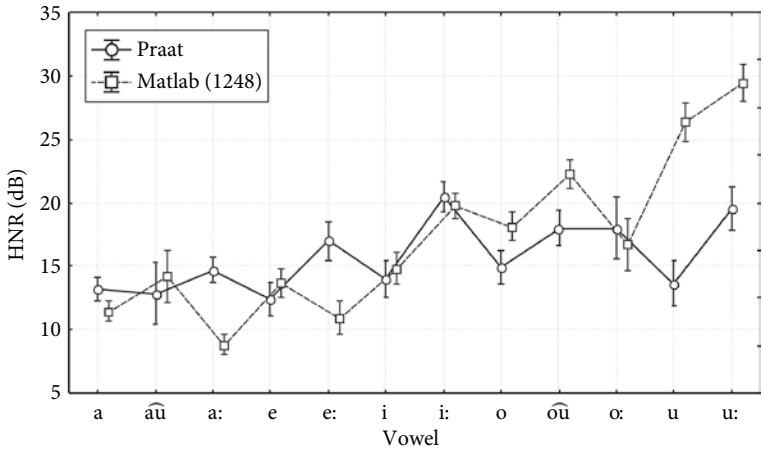

Figure 3. The results of ANOVA analysis for the AC-based and wavelet-based HNR estimates.

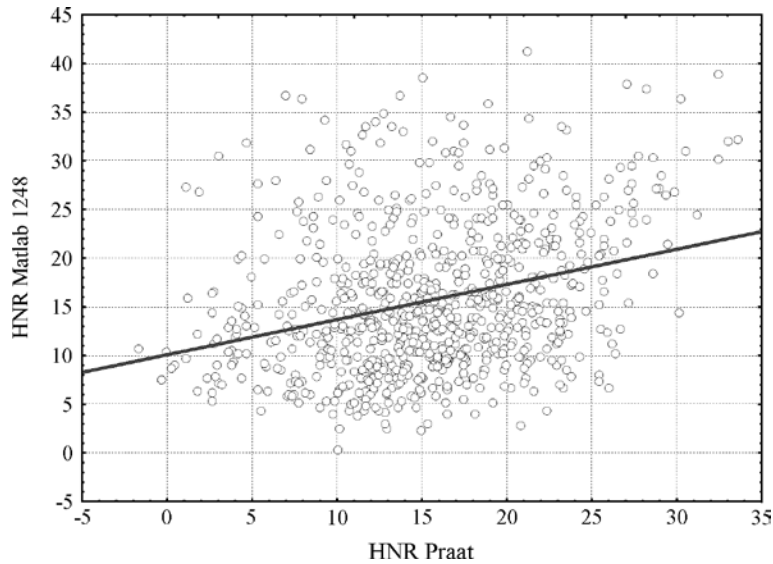

Figure 4. Correlation between the AC-based and wavelet-based HNR estimates. N = 791, $p < 0.05$.

method ($F(11,779) = 108.49$; $p < 0.001$). However, we observed a noticeable difference between the AC- and wavelet-based values (mean absolute difference of 6.34 dB). The correlation between the AC- and wavelet-based estimates proved to be weak $r = 0.3$, $p < 0.05$ (see Figure 4).

4. Conclusions

The aim of this paper was to compare the Fourier and the wavelet transforms on the example of harmonicity. The Fourier transform is commonly used in speech analysis. The tools to retrieve the information about acoustic cues from speech signal have been implemented into graphical user interfaces that are easy to use also for non-technicians. Praat is one of these examples. To measure HNR in Praat we need to use a couple of buttons. Therefore we chose the AC-based HNR values measured in Praat as reference values for our comparison. In contrast to Praat, MATLAB is a general computing environment. It is not tailored for speech processing. In the wavelet toolbox we find tools to analyse the signal, but there are no direct options to detect F0 or to measure HNR. Therefore we had to propose an algorithm in MATLAB to obtain wavelet-based HNR values. We expected that this method would provide similar results as Praat. This hypothesis, however, proved to be wrong. There was no significant correlation between both data sets detected. We observed significant differences between the AC-based and wavelet-based values. We put these findings down to the method of HNR estimation in MATLAB. In our approach the signal was decomposed, a specific frequency band was chosen to represent the periodic component, and HNR was computed. We believe that the reason for the obtained discrepancy in the results is the fact that our wavelet-based approach counts only with additive noise as a negative factor influencing the HNR value and it does not reflect F0 perturbations. This became especially obvious for back vowels. The energy of the first and the second formant is in the case of back vowels concentrated below 1200 Hz (Palková, 1997: 172–175). Figure 5 shows an example of a short vowel [u] with AC-based HNR value 13.8 dB measured in Praat. For the same example the wavelet-based approach computed HNR value of 33.65 dB. Such a high value is in accordance with the energy distribution in the lower frequency bands (see Figure 2). It is obvious that the waveform decays in time. This, however, seems not to be reflected in the wavelet-based method.

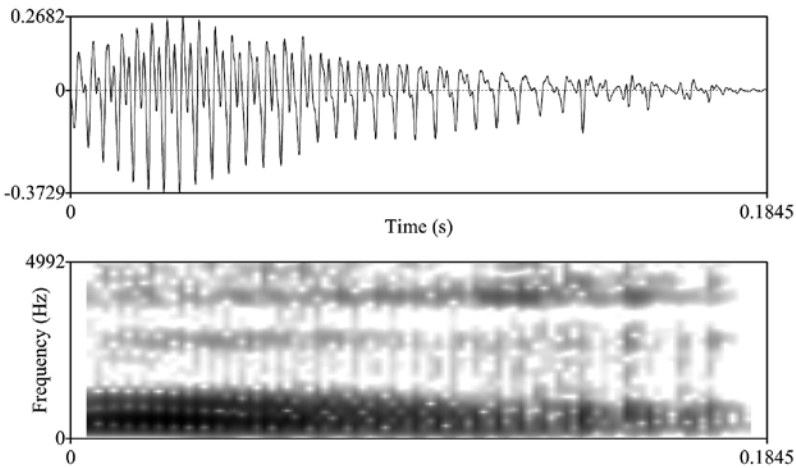

Figure 5. Oscillogram and spectrogram of vowel [u] with 13.18 dB HNR value calculated in Praat using AC-based approach and 33.65 dB HNR calculated in MATLAB using wavelet-based approach.

On one hand, Ferrand (2002) indeed informs that the harmonicity measure quantifies the amount of additive noise in the voice signal. On the other hand, Murphy (2006) reports that the harmonicity is influenced not only by the additive noise but also by inter-period glottal waveshape differences and F0 perturbations. Our results proved that the employed methods for harmonicity measurement are not comparable and therefore our results are inconclusive with respect to the comparison of the Fourier and the wavelet transforms and their advantages and disadvantages for speech analysis. We assume that to reach the correlation between the AC- and wavelet-based HNR values we would also need to take into consideration F0 perturbations.

A possibility for future research could be an implementation of HNR based on the Fourier transform as used in Praat into MATLAB environment and their comparison. Also an analysis of a non-stationary signal where a wavelet-based approach could prove more convenient could be interesting with regard to the confrontation of the wavelet- and DFT-based analysis of speech. As a suitable material for such research the transitions between vowels and voiceless consonants appear. Such an analysis for AC-based HNR estimation was performed in Praat with promising results (see Heranová and Skarnitzl, 2011).

ACKNOWLEDGEMENTS

This research was supported by an internal grant VG134 of Faculty of Arts of Charles University in Prague.

REFERENCES

Awan, S. N. & Frenkel, M. L. (1994). Improvements in Estimating the Harmonics-to-Noise Ratio of the Voice. Journal of Voice, 8/3, pp. 255–262.

Boersma, P. (1993). Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound. IFA Proceedings, 17, pp. 97–110.

Boersma, P. & Weenink, D. (2011). Praat: doing phonetics by computer [Computer program]. Version 5.2.15, retrieved on February 12, 2011 from <http://www.praat.org>.

Cox, N. B., Ito, M. R. & Morrison, M. D. (1989). Technical Considerations in Computation of Spectral Harmonics-to-Noise Ratios for Sustained Vowels. Journal of Speech and Hearing Research, 32, pp. 203–218.

Ferrand, C. T. (2002): Harmonics-to-Noise Ratio: An Index of Vocal Aging. Journal of Voice, 16/4, pp. 480–486.

Heranová, J. & Skarnitzl, R. (2011). Využití harmonicity při fonetické segmentaci řeči. Akustické listy, 17/4, pp. 3–9.

Johnson, K. (2003). Acoustic and Auditory Phonetics. Oxford: Blackwell Publishing.

Ladefoged, P. (1996). Elements of acoustic phonetics. Chicago: The University of Chicago press.

Mallat, S. (1998). A wavelet tour of signal processing. San Diego: Academic Press.

Mathworks (2012). MATLAB R2012a. (Student version 7.14.0.739).

Murphy, P. J. (2006). Periodicity Estimation in Synthesized Phonation Signals Using Cepstral Rahmonic Peaks. Speech Communication, 48, pp. 1704–1713.

Palková, Z. (1997). Fonetika a fonologie češtiny. Praha: Karolinum.

Qi, Y. & Hillman, R.E. (1997). Temporal and Spectral Estimations of Harmonics-to-Noise Ratio in Human Voice Signals. Journal of the Acoustical Society of America, 102/1, pp. 537–543.

Selesnick, I. W. (2007). Wavelet Transforms – A Quick Study. Retrieved on 25 November 2012 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.139.148.
Yumoto, E., Gould, W. J. & Baer, T. (1982). Harmonics-to-Noise Ratio as an Index of the Degree of Hoarseness. Journal of the Acoustical Society of America, 71/6, pp. 1544–1549.
Zhao, S., Bu, F., Sun, Y. & Han, C. (2003). Study on HNR in transmitted sound signals. In: Proceedings of the 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering. Beijing, pp. 580–584.
Zimmermann, J. (2002). Spektrografická a škálografická analýza akustického rečového signálu. Prešov: Náuka.

FOURIEROVA A VLNKOVÁ TRANSFORMACE V PROCESU ŘEČI: PŘÍPAD HARMONICITY

Resumé

Tento příspěvek se věnuje problematice analýzy řečového signálu, konkrétně Fourierově a vlnkové transformaci. Nejčastějším a nejrozšířenějším způsobem pro získání frekvenční reprezentace signálu je ve fonetickém výzkumu Fourierova transformace. Ta má však i své nevýhody, přičemž tou hlavní je nepřímý vztah mezi časovým a frekvenčním rozlišením. V porovnání s Fourierovou transformací představuje vlnková transformace poměrně novou a pro účely fonetického výzkumu neprozkoumanou metodu, která, jak se zdá, umožňuje přirozenější rozklad signálu než klasická transformace Fourierova. Cílem tohoto příspěvku bylo srovnání obou metod analýzy signálu na příkladu harmonicity, poměru harmonických a šumových složek signálu v decibelech. Pro srovnání jsme jako referenční hodnoty použili hodnoty harmonicity naměřené v programu Praat, volně šiřitelném programu k akustické analýze a zpracování zvuku, který ke zjištění periodicity signálu využívá metodu autokorelace. V prostředí MATLAB jsme poté navrhli metodu měření harmonicity, která využívá k rozkladu signálu vlnkovou paketovou analýzu (anglicky wavelet packet analysis). Srovnáním hodnot harmonicity získaných pomocí obou metod jsme dospěli k závěru, že námi navržená metoda v prostředí MATLAB nezohledňuje perturbace F0, které do signálu také zanášejí šum, a tudíž se nepotvrdila naše hypotéza, že výsledky obou metod budou vykazovat korelaci.