INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Analytic Filter Bank for Speech Analysis, Feature Extraction and Perceptual Studies

Unto K. Laine

Aalto University, Espoo, Finland
unto.laine(at)aalto.fi

Abstract

Speech signal consists of events in time and frequency, and therefore its analysis with high-resolution time-frequency tools is often of importance. Analytic filter bank provides a simple, fast, and flexible method to construct time-frequency representations of signals. Its parameters can be easily adapted to different situations from uniform to any auditory frequency scale, or even to a focused resolution. Since the Hilbert magnitude values of the channels are obtained at every sample, it provides a practical tool for a high-resolution time-frequency analysis.

The present study describes the basic theory of analytic filters and tests their main properties. Applications of analytic filter bank to different speech analysis tasks including pitch period estimation and pitch synchronous analysis of formant frequencies and bandwidths are demonstrated. In addition, a new feature vector called group delay vector is introduced. It is shown that this representation provides comparable, or even better results, than those obtained by spectral magnitude feature vectors in the analysis and classification of vowels. The implications of this observation are discussed also from the speech perception point of view.

Index Terms: speech analysis, pitch-synchronous analysis, time-frequency methods

1. Introduction

Speech signals have unique properties in both time and frequency domains. The quasi-periodic glottal excitation is the main source of the temporal structures in voiced phones. The primary excitation of the vocal tract resonator occurs at the glottal closure and the secondary at its opening. If the closure is tight it causes an almost impulse like excitation. Whereas, if glottis is leaking the excitation consists of a pulse-noise combination and the resonances of the vocal tract have broader bandwidths. The excitation at the glottal opening is a wide-band noise. The glottal opening also changes the acoustic load of the tract, causing leakage of the acoustic energy from the vocal tract to the subglottal cavities. This causes extra attenuation of the resonances of the tract before the next glottal closure, leading to pitch synchronous modulation of the formant bandwidths. All these factors depend on speaker and speaking style [1].

The formant bandwidths are also affected by the lip radiation impedance. The second and higher formants are especially sensitive to the impedance matching between the vocal tract and the lip radiation load. Open vowels have small radiation impedance causing extra attenuation of the formants. In the closed vowels the situation is opposite. The impedance matching also depends on the vocal tract shape. Thus the formant Q-values (Q=Fx/B) are changing dynamically approximately between 5–25 during continuous speech. These changes have an important role in speech quality because they directly affect the formant amplitude values.

Combined pulse and noise excitations appears also, e.g. in the release phases of the stop consonants. The detailed time-frequency structures of these short bursts depend on the place of articulation. In these cases important information is packed in a short time window, and these events are easily lost when a typical time-frequency analysis using windows across several tens of milliseconds is applied. It is obvious that studies related to these details of speech signals need an efficient method providing a high time-frequency (TF) resolution.

Up to now the group delay (GD) and phase spectrum of speech has been studied almost entirely by the short-term Fourier transform [2, 3, 4, 5]. Some of the proposed methods are applicable for minimum phase signals only and speech is not such. Also, estimation of GD with Fourier method is computationally more demanding and noise sensitive than to obtain the magnitude spectrum. In this work the GD is observed directly from the TF-distribution without the usage of any complex phase based computations.

The paper starts with a short introduction to the analytic signals and filters. Because the analytic filters provide signal Hilbert envelope values at each sample, they support a high temporal resolution signal analysis. An efficient realization of an analytic filter bank (AFB) is described and its performance tested by compact Gabor pulses. The usage of AFB in pitch synchronous analysis of speech is described. This part of the study describes a new method to estimate the GD that provides information on the formant frequencies and bandwidths. In a preliminary test where single pitch periods of vowels are classified the error probability obtained by the GD feature is less than one third of that of the classical STFT based log-magnitude feature.

2. Analytic signals and filters

A complex valued signal where its real and imaginary parts form a Hilbert pair is called an analytic signal. An analytic filter takes in real valued samples and produces an analytic signal from them. When dealing with a simple resonator (in classical terms second order system) we need two real valued coefficients to control its resonance frequency and bandwidth. However, the same result can be obtained with an analytic filter with one complex valued coefficient.

2.1 Analytic resonator

The transfer function of an analytic resonator is given by Eq. (1), where the parameter r controls its bandwidth and $\omega_0 = 2\,\pi\,f_0$ its center frequency. The parameter r is given by

http://dx.doi.org/10.21437/Interspeech.2017-1232

r=exp[-π*B/f_S], where B is the bandwidth in Hz and f_S the sampling frequency.

$$H(z) = \frac{1}{1 - z_0 z^{-1}}, z_0 = re^{j\omega_0 / f_S} \quad (1)$$

An example of an analytic resonator and its impulse responses y[n] is given in the Figure 1 with real (blue) and imaginary (red) components. The envelope of the output signal is easily obtained by abs(y) which is a smooth exponential function. Thus the analytic resonator is able to give the instantaneous amplitude value at every sample. Whereas in the case of a classical second order resonator the abs-value of the output must be filtered before the amplitude estimate, which reduces the temporal resolution considerably in comparison to the AFB [2, 6]. Input $d[n]$ is a unit impulse. In this paper the focus is in the magnitude envelope of these filters.

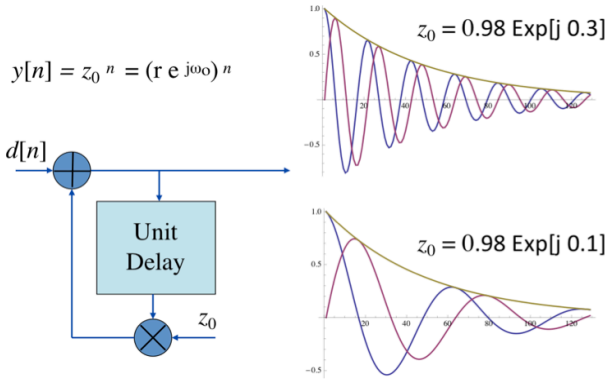

Figure 1: *Example of an analytic resonator. Blue curves represent real parts of the output, red and yellow curves imaginary part and abs-values correspondingly.*

2.2 Time-frequency resolution test

A Gabor pulse defined by Eq. (2) is compact in time and frequency [7]. Now it is applied for testing of the AFB.

$$h(t) = \exp(-\alpha^2 t^2)\sin(2\pi f t) \quad (2)$$

where f is its center frequency and α controls its bandwidth. Figure 2 represents the time-frequency distribution of the Gabor pulse with center frequency of 4 kHz analyzed with AFB constructed with 160 filters distributed uniformly over 8 kHz band and each having bandwidth of 50Hz.

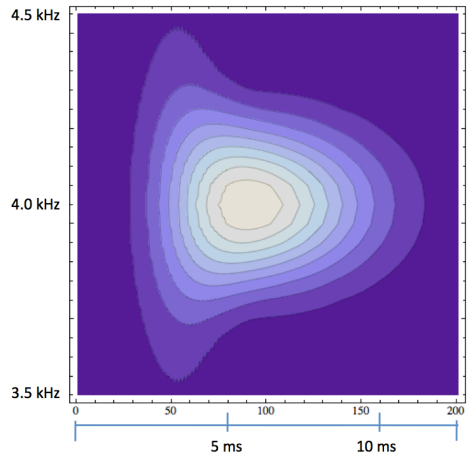

Figure 2: *Magnitude envelope of the Gabor pulse of 4 kHz analyzed by an AFB. x: Time in ms and samples, y: Channel center frequencies.*

With an optimal α parameter value the detected Gabor pulse (in Figure 2) had Heisenberg time-frequency resolution Δt Δ ω = η around 1.2 (see [7], pp. 86-90). The theoretical limit of η of a linear causal system is 0.5. However, the vocal tract resonances have η around three or even larger as will be shown by the following analysis. The value of η as well as the ratio Δt /Δ ω can be controlled by convolving the output spectral vectors of the AFB by an analytic window. A window of five points: 0.08 + j 0.582674, 0.54 + j 1.3454, 1. 0, 0.54 – j 1.3454, and 0.08 – j 0.582674 was applied in this study. This window was constructed experimentally, due to the lack of rigorous theory at present in this field. The selected window provides a resolution (η =1.2) and proves to be a good choice for vowel analysis. An example of the time-frequency distribution of the Finnish [e] sound is given in the Figure 4. The center frequency of the channel number 100 is 5 kHz.

2.3 η-values of acoustic resonator

The selection of an optimal η–value for speech analysis needs knowledge of the time-frequency properties of a typical vocal tract resonator [8]. A simulation was made to study this aspect. The resonator can be modeled with

$$H(j\omega) = \frac{\omega_0^2}{\omega_0^2 + j2\omega\varsigma\omega_0 + (j\omega)^2} \quad (3)$$

where w_0 is its center frequency and ζ its attenuation constant that is also related to its Q-value by Q = 1/ (2 ζ). The impulse response or this resonator is similar to the real components shown in the Figure 1. The time-frequency properties of the impulse responses were analyzed in terms of the η–value when the ζ–parameter was varied (Figure 4).

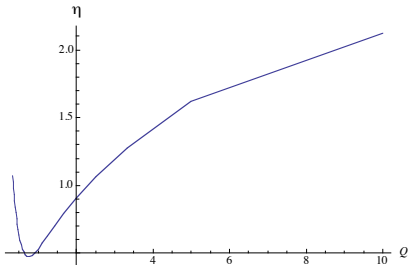

Figure 3: *Resonator η–value as a function of its Q-value.*

Because the most common resonance Q-values observed in real speech are above 6 and their η–values larger than 1.5, the parameters selected for the AFB fit the needs of formant analysis well.

2.4 Group delay of acoustic resonator

During the AFB experimentation it was noted that the energy at those channels where the VT resonances are located was delayed. The amount of delay was proportional to the Q-value of the resonance. A potential explanation is that the AFB is able to estimate the formant group delays present in real speech. In order to verify this the theory of the GD is described next.

The group delay of the resonator in Eq. (3) is defined by

$$D = -\frac{d\varphi(\omega)}{d\omega}, \quad (4)$$

where $\varphi(\omega)$ denotes the phase of the $H(j\omega)$. The group delay

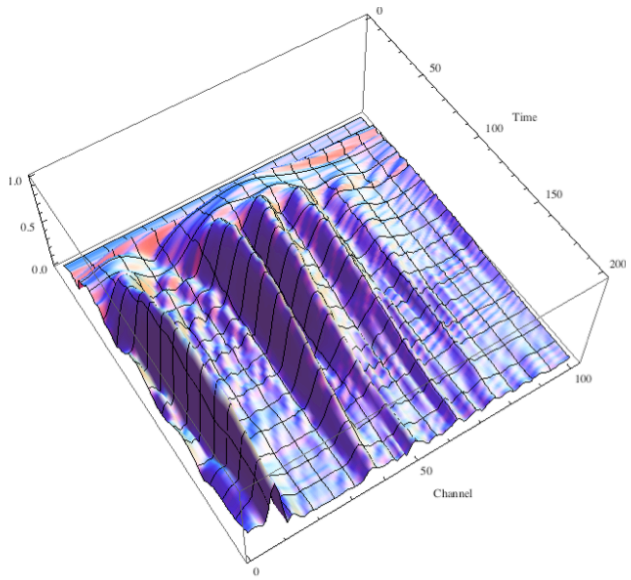

Figure 4: *Time-frequency distribution (magnitude) of one pitch period of the Finnish* [*e*] *sound. Frequency band: 0-5 kHz and time segment: 10 ms (=200 samples, f_S=20 kHz).*

of a resonator with the center frequency of 1 kHz and bandwidth of 100 Hz is depicted in Figure 5. The peak value of the group delay is about 3 ms. The delay increases when the bandwidth gets smaller. The observation that the AFB is able to estimate the format group delays led to the next question; is it possible to create a new type of feature vector based on this information and how useful is this feature in comparison to spectral magnitude based features?

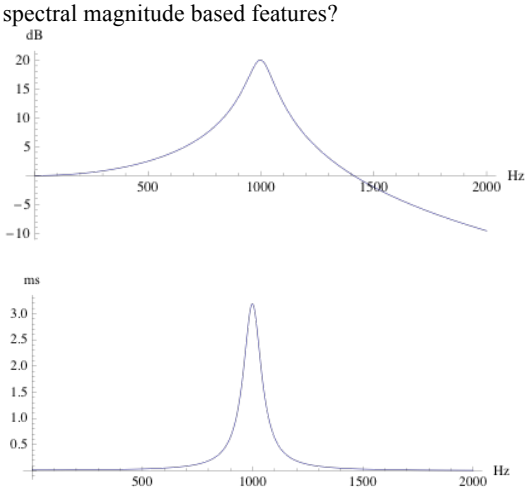

Figure 5: *Magnitude and group delay of a second order resonator with bandwidth of 100 Hz.*

3. Pitch periods and group delay

3.1 TF-structures of vowel pitch periods

Before studying the group delay aspect, the general TF-structures of pitch periods of the eight Finnish vowels were analyzed in order to study their temporal energy distributions and to find out the time instances (inside pitch period) where the most distinctive TF-structures are located. Thirty-two periods used in this test were selected from long vowels pronounced in isolation by a male speaker. The envelopes in the Figure 1 do not indicate the group delays in question. However, when the impulse response is fed to a filter bank the envelope of the channel close to the resonance frequency has a larger lag in comparison to the other channels. According to the group delay theory, this delay should depend on the actual *Q*-value of the formant. Figure 4 reveals that 1.0–1.5 ms after the glottal closure the spectrum is almost flat indicating the strong excitation (note that 20 samples is 1 ms in this analysis). Naturally, this part of the TF-distribution of different vowels is very similar and therefore cannot contribute much to their discrimination. The end part of the pitch period, where the glottis is open, is often noisy and the formants are strongly attenuated. Therefore, the TF-segment providing the best contrast for discrimination should be found somewhere in the middle of the pitch period. Note that in the Figure 4 the strongest formant amplitudes are found about 3 ms after the glottal closure.

The pitch period energy distribution over time is given by the time-marginal of the TF-distribution. The mean of the time-marginals of the eight vowels is shown by the blue curve in Figure 6. The red curve represents the mean of the *spectral contrast* distributions of the vowels. This measure goes to zero when all elements of the spectral vector have the same value and increases when it has high maxima and deep valleys. This contrast measure is simply the length of the spectral envelope curve (total sum of the abs-differences of the adjacent elements). These two curves indicate, that the energy distribution activates about 2 ms (40 samples) before the best spectral contrast. The next question is which of these two aspects are more important when looking after the highly discriminative areas in the TF-distributions of the vowels.

The locations of the highly discriminative spectral vectors in the TF-distributions of the vowels were studied by picking up adjacent spectral vectors with a sliding window of varying size and computing the Euclidean distances between the obtained TF-matrices and selecting the value of the smallest distance at each window position (the worst case). Finally, the window position where these minimum distances reached their maximum was stored. When this procedure was repeated with different window sizes it was possible to localize the spectral vectors with the highest discriminative property. The rectangular pulses in Figure 6 represent the optimal window positions. Their amplitudes are related to the found maximum (of the smallest Euclidean distances) normalized by the actual window length. The result clearly indicates that the best discrimination is not achieved around the energy maximum but closer to the maximal spectral contrast about 3 ms later.

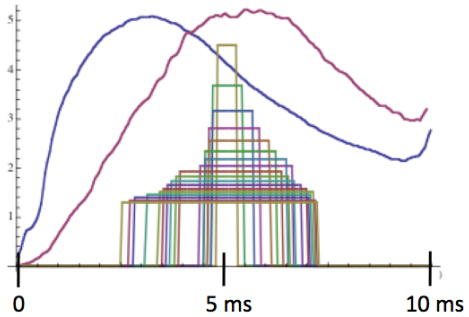

Figure 6: *Mean energy (blue) and spectral contrast (red) distribution over pitch periods of the eight synchronized Finnish vowels. Rectangular boxes: locations and sizes of the windows within the pitch periods where the discrimination reached its maximum. x: time in ms. y-axes: arbitrary scales.*

3.2 Pitch analysis

The mean of the AFB channel envelopes produce a smooth pulsation for voiced sounds that can be relatively easily applied to pitch period estimation, e.g. by looking for the locations of the maximum derivative of the mean envelope (see Figure 7).

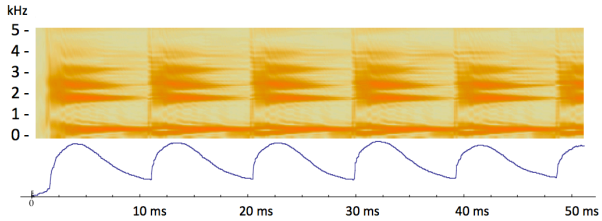

Figure 7: *AFB spectrogram of the Finnish* [*e*] *vowel with the mean envelope of the channels.*

3.3 Group delay – a new feature vector

The closer examination of the TF-properties of the AFB magnitude revealed that the channels where formants are located reach maximum amplitude later than the others. This means that the AFB method is able to estimate the GDs directly from its magnitude information without any complex phase related processing. Thus the AFB provides a new *group delay feature* that is next compared to more classical features.

The estimation of the GD at any AFB channel is relatively simple. First, the channel magnitude $m(t)$ is smoothed and its sum is normalized to one. Then a dot product $t_m = n(t).m(t)$ is computed, where $n(t)$ is a range of integers from one to the number of samples in m. This provides an estimate of the time instant t_m of center of $m(t)$. The group delay is estimated by $t_m - t_0$, where the t_0 denotes the time instance of the glottal closure.

Three feature vectors were compared; log-magnitude vector of each pitch period (Hamming window & STFT), log-magnitude vector of the AFB TF-distribution frequency marginal, and the group delay vector. Each set has 32 vectors, four for each vowel. Together 448 cross-correlations are computed for each feature type between the vowel classes and 48 within the classes. Their histograms were modeled by Gaussian distributions and finally the probabilities for miss and false-alarm cases solved (see Figure 8).

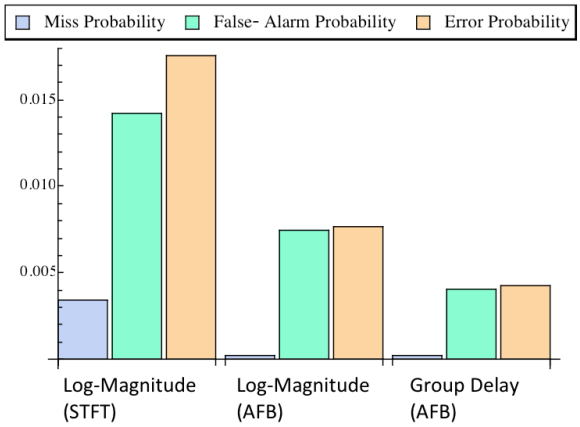

Figure 8: *Error probabilities in vowel classification (one pitch period) with three different feature vector types.*

Both AFB based features performed clearly better than the classical STFT-based log-magnitude feature. This may be due to the filter bank design were the parameters were optimized for VT resonances.

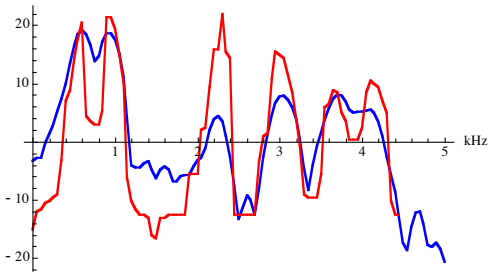

Figure 9: *Comparison of log-magnitude (blue) and GD-feature (red) of the vowel* [*a*]. *x-scale 0-5 kHz, y-scale is arbitrary.*

Figure 9 shows a comparison of log-magnitude of the mean of 65 spectral vectors with the best discrimination ability to the GD-feature vector of the same pitch period. The difference in quality of the features is clearly visible.

4. Discussion and conclusion

During the last decades psychoacoustic research has collected evidence of the extremely high temporal resolution of the human auditory system. Even one micro second time differences are perceived in binaural directional hearing and also in monaural listening the cross-channel synchrony effects are of importance [9, 10, 11]. Based on this knowledge it is likely that the details in group delays discovered by the AFB method may have a significant perceptual value: those auditory channels where the formants are located will get the excitation later than the other channels that are synchronized to the main glottal excitation. The AFB method provides a practical tool to start a closer examination of this phenomenon.

This study focused entirely on the magnitude processing of the AFB output. However, its phase processing may provide even more possibilities, e.g., to analyze the dominating instantaneous frequencies at the channels.

Historically, Fourier based phase and group delay processing have proven to be a computationally demanding process especially when real speech signals are to be processed. This is because they are not minimum phase signals. The group delay aspect of the new AFB design came up as a surprise after the careful design of the method. In this (first) paper of the AFP method the focus was in its basic theory and applications to pitch synchronous vowel analysis. The obtained results motivate the future development and application of the method more widely in speech processing.

This work introduced the AFB method and showed its performance in pitch synchronous analysis. It was shown that the most important and distinctive structural information is located in the middle part of the pitch periods. The most promising new outcome was the group delay feature vector that models vowels more orthogonally than the classical log-magnitude features.

5. Acknowledgements

The comments and suggestions of Okko Räsänen. Shreyas Seshadri and four unknown reviewers are kindly acknowledged.

6. References

[1] J. L. Flanagan, *Speech analysis synthesis and perception*, Springer-Verlag, Berlin Heidelberg, 444p., 1972.

[2] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Prentice–Hall, 1975.

[3] A. V: Oppenheim, J. S. Lim, "The importance of phase in signals", Proc. IEEE 69, pp. 529-541, 1981.

[4] A. Stark and K. Paliwal, "Group-Delay-Deviation Based Spectral Analysis of Speech," *Interspeech 2008, Brighton*, pp. 1083-1086, 2008.

[5] P. Mowlace, R. Saeidi, Y. Stylianou, "Advances in phase-aware signal processing in speech communication", *Speech Communication*, **81**, pp.1-29, 2016.

[6] C. M. Vikram, S. R. Mahadeva Prasanna, "Epoch Extraction From Telephone Quality Speech Using Single Pole Filter", IEEE/ACM Trans. Audio, Speech and Signal Processing, **25**, No. 3, 2017.

[7] L. Cohen, *Time-frequency analysis*, New Jersey, Prentice Hall, 1995.

[8] G. Fant, *Acoustic theory of speech production,* Mouton, The Hague, 1970.

[9] J. O. Pickles, *An Introduction to the Physiology of Hearing,* 430p., Brill, 2013.

[10] K. Krumbholz, R. D. Patterson, A. Nobbe, and H. Fastl, "Microsecond temporal resolution in monaural hearing without spectral cues?" *JASA*, **113**, 5, 2790, 2003.

[11] K. Krumbholz &al., "The effect of cross-channel synchrony on the perception of temporal regularity", *JASA*, **118**, 946, 2005.