ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SEPARATION OF SPEECH SIGNALS USING ITERATIVE MULTI-PITCH ANALYSIS AND PREDICTION

Matti Karjalainen and Tero Tolonen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O.Box 3000, FIN-02015 HUT, Finland
matti.karjalainen@hut.fi, tero.tolonen@hut.fi

10.21437/Eurospeech.1999-484

ABSTRACT

A model for multi-pitch analysis is extended into an iterative multi-pitch analysis and prediction (IMPAP) scheme. The method is efficient in finding harmonic complex tones, such as voiced speech signals, in a mixture of such signals and possible noise background. It can also be used to separate the signal into perceptually relevant speech components. The method may be used in applications ranging from speech transmission (enhancement) to recognition (noise and extra sound rejection).

1. INTRODUCTION AND MOTIVATION

The problem of separation of multiple speakers and other sound sources in a single sound signal is often encountered in speech applications ranging from speech transmission (enhancement) to recognition (noise and extra sound rejection). While perfect separation of constituent signal components appears generally impossible at this time, reasonable results that are useful in practical applications may be accessible, as demonstrated in the examples below. In particular, separation of voiced speech components may be performed with reasonable accuracy.

Typically, separation of voiced components is based on analysis of the fundamental frequencies or, preferably, pitches. Many principles have been proposed for the modeling of human pitch perception and for practical pitch determination of audio or speech signals [1, 2]. Recently it has been demonstrated that a peripheral auditory model which uses time-domain processing of periodicity properties shows ability to simulate many known features of pitch perception which are often considered to be more central [3, 4]. Time-domain models are attractive since auditory processes may be simulated with relatively straightforward DSP algorithms. Additional features may be readily included using, e.g., frequency domain algorithms, if desired. A practical problem with most of the "time-domain" models is, however, their computational requirements. They typically include an auditory filterbank with 32–128 channels, which makes it impossible to operate such a model in real time in a typical desktop computer.

Recently, we have reported a multi-pitch and periodicity analysis model that is computationally relatively efficient and still auditorily motivated [5]. In this paper, we apply that model for pitch analysis and further extend it into an iterative multi-pitch analysis and prediction (IMPAP) scheme. The method is efficient in finding harmonic complex tones, such as voiced speech signals, in a mixture of such signals and possible noise background. It can also be used to separate the signal into perceptually relevant speech components.

A similar approach has been recently presented for identification of vowels based on template matching and iterative identification [6]. It is based on a multi-channel pitch analysis model that can also be used in segregation of concurrent voiced sounds, e.g., voiced speech [7].

2. REDUCED COMPLEXITY MODEL OF AUDITORY PITCH PERCEPTION

The Meddis-O'Mard unitary pitch perception model is a representative example of recent 'time-domain' pitch analysis models. A key component is a gammatone filterbank [8] which is used to simulate the frequency selectivity of the peripheral hearing. The signal is split into channels such as ERB (equivalent rectangular bandwidth) channels and each channel is half-wave rectified and lowpass filtered (about 1 kHz) in order to simulate the activity of the hair cells. Signal periodicity is next extracted in each channel by computing its autocorrelation function (ACF) or a similar periodicity measure. Finally, the ACFs are summed from each channel to yield a summary autocorrelation (SACF) that shows the overall periodicity properties of the incoming signal. For more details, see [3, 4].

Meddis and O'Mard have compared the behavior of the unitary pitch perception model with results from psychoacoustical experiments and shown that the model is capable of simulating several important or interesting special cases of perception, at least qualitatively [3].

Our simplified pitch analysis model is illustrated in Fig. 1. The middle part of the model ($x_1 \rightarrow x_2$) corresponds to the functionality of the Meddis-O'Mard model. The simplicity of the model is based on the division of the audio frequency range to only two subchannels. Low frequencies below 1 kHz are analyzed directly by autocorrelation while high frequencies above 1 kHz are first (half-wave) rectified and low-pass filtered and then the autocorrelation is computed. Summary autocorrelation is now the sum from only two subchannels. This approach, in comparison to the multi-channel models, is based on assumption that at low frequencies for this particular task the auditory system essentially acts as a simple linear channel before the periodicity detector and that above 1 kHz it does envelope following and then similar periodicity detection of the envelope. This two-channel analyzer is naturally much more efficient computationally than a multi-channel pitch analyzer.

Several details of the model in Fig. 1 are important for proper functioning. The 'Pre-whitening' box in the model is a frequency-warped version of linear prediction (WLP, order 12 for sample rate of f_s = 22 kHz), as described in [9]. This effect may be considered somewhat similar to the adaptation in hair cell models. The block 'Periodicity detection' is based on the autocorrelation function. In the model it is computed through the discrete Fourier transform (DFT) and its inverse (IDFT) as $corr(\tau) = \mathrm{IDFT}\{|\mathrm{DFT}\{x(\tau)\}|^k\}$ where exponent $k = 2/3$ instead of $k = 2$ for normal autocorrelation and τ is the time lag variable.

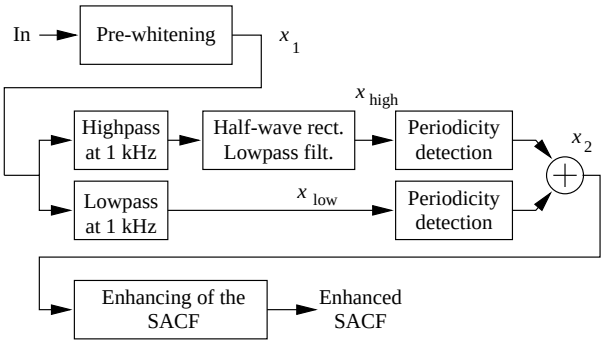

Figure 1: A block diagram of the proposed model.

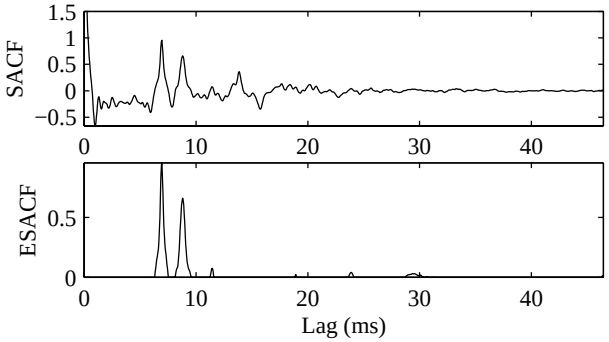

Figure 2: Example of enhancing the SACF representation.

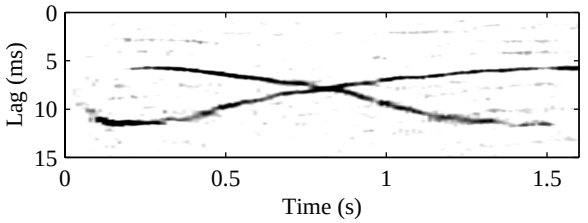

Figure 3: Example of pitch analysis (lag vs. time map) of a signal consisting of two similar vowels with time-varying pitches.

In the frequency domain this operation shows resemblance to the loudness scaling of magnitude [10]. One more detail of implementation is that actually the two (second order) 1 kHz lowpass filters in the middle part of Fig. 1 model include a highpass property, e.g. with a cutoff at 80 Hz, in order to remove the down-ramp baseline of the correlation function up from zero time lag. A Hamming window of 46.4 ms (frame size 1024 samples for $f_s = 22$ kHz), yielding about 25 ms effective window length, is used. The hop size of SACF computation window in the model is 10 ms. These implementation features have been selected based on experimentation with various practical signals, see [5] for details.

The peaks in the SACF curve produced as x_2 output of the model in Fig. 1 are relatively good indicators of potential pitch periods in the signal being analyzed. Such a summary periodicity function contains, however, much redundant and spurious information that makes it difficult to estimate which peaks are true pitch peaks. The autocorrelation function generates peaks at all integer multiples of the fundamental period. Furthermore, in case of musical chords the root tone, the common periodicity, often appears very strong though in most cases it should not be considered as a fundamental period of any source sound. To be more selective, a peak pruning technique similar to [11] is used in the model.

The technique is the following. The original SACF curve, as demonstrated above, is first clipped to positive values and then time-scaled (expanded in time) by a factor of two and subtracted from the original clipped SACF function, and again the result is clipped to have positive values only. This removes all repetitive peaks with double the time lag where the basic peak is higher than the duplicate. This also removes the near zero time lag part of the SACF curve. This operation can be repeated for time lag scaling with factors of three, four, five, etc., as far as desired, in order to remove higher multiplies of each peak. The resulting function is called here the enhanced summary autocorrelation (ESACF).

Figure 2 illustrates the enhancing of the SACF representation. The top plot presents an SACF of a mixture of two Finnish vowels `/ae/` spoken by a male. The bottom plot depicts the corresponding ESACF representation. It is obvious that the two peaks corresponding to pitch periods are more clearly identifiable at the ESACF.

An example of pitch detection is shown in Fig. 3 which illustrates the temporal evolution of the enhanced SACF function in the analysis of a mixture signal containing two simultaneous vowel pitch glides. Both vowels are Finnish `/ae/` sounds mixed from the same speaker, one gliding in pitch from low to high and the other from high to low pitch. The pitch analysis information is shown as a spectrogram-like presentation which clearly indicates the fundamental periodicity trajectories. Some spurious peaks appear at low level but no systematic high level phantom trajectories are found in this relatively easy case where the two vowels have about the same amplitude.

3. ITERATIVE DETECTION, PREDICTION, AND SEPARATION

In the previous example the two pitch trajectories were easily separable. In many cases, however, prominent components mask the weaker ones in the ESACF representation, particularly if there are several voiced components present. The ESACF estimation is typically relatively accurate for the most prominent components but with weaker components, even if they are identifiable in the ESACF representation, the peak location may give an inaccurate value for the pitch. A natural idea is to iteratively identify and separate the prominent components in the signal so that the weaker components may be identified with improved reliability.

Figure 4 shows a block diagram of the iteration principle. In each iteration, the pitch analysis described in the previous section provides the ESACF representation. A prominent component is identified in the ESACF and its pitch is derived from the location of the corresponding peak. The contribution of the most prominent component is separated in the signal, and the pitch analysis is performed again. The iteration is continued until all significant components have been identified.

The pitch analysis block of Figure 4 corresponds to the functionality of the pitch analysis model in Figure 1 and provides the ESACF representation as its output. The identification of the most prominent component is performed by detecting the maximum value of the ESACF representation. In some applications, however, it may be of advantage to use a different approach. For instance, an event in the sound signal, such as a musical tone, may be identified by analysis of consecutive ESACF frames. In this case, even when the tone may not produce the highest peak in all the ESACF frames, its pitch may be identified based on rules of pitch continuation and the component may be separated in the original signal. An example of this approach has been pre-

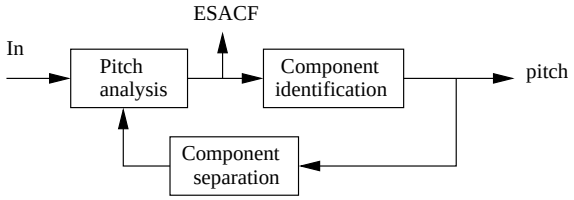

Figure 4: A block diagram of iterative identification and separation of voiced sounds.

sented in [12] where an excerpt of a musical melody was separated from accompaniment using the ESACF representation for identification of the note events.

If harmonic tones are closely spaced in fundamental frequency and particularly if one of the tones is considerably weaker than the other, they may result in a merged peak in the ESACF representation. An estimate of better accuracy may then be derived using other methods for fine-tuning the initial estimate obtained in the ESACF representation. One such approach is to use *nonlinear least-squares* (NLS) estimation of the fundamental frequency. This method is a straightforward modification of a similar parametric estimation method for sinusoids (see, e.g., [13]). The fundamental frequency is determined by minimizing the cost function

$$G(f_0) = \sum_{n=0}^{N-1} \left| y(n) - \sum_{k=1}^{N_{\mathrm{harm}}} e^{i(2\pi k f_0 n)} \right|^2 \qquad (1)$$

over a meaningful range of the fundamental frequency f_0. Parameters N and N_{harm} are the window length and the number of harmonics, respectively. The following example shows how the (NLS) fundamental frequency estimation performs with practical signals.

Note that above we assumed that the signal is strictly periodic, i.e, the frequencies of harmonics are in exactly integer ratios. With relatively stationary voiced speech signals, the method typically works well. If the NLS method is applied, e.g., to string instrument sounds that exhibit dispersion, it is straightforward to modify the NLS model to include optimization of the dispersion parameter.

There are several approaches for separation of the detected components. One approach is to use sinusoidal modeling [14] in which the frequencies, amplitudes, and phases of the sinusoids corresponding to a voiced signal component are detected and used in synthesis and subtraction of the detected component. Sinusoidal modeling separation based on the ESACF representation is discussed in [12]. In this contribution, we apply linear filtering to remove the detected signal component. Note that while conceptually sinusoidal modeling is more natural for separation of additive signal components, linear filtering is computationally superior and produces satisfactory results in most cases.

The pitch predictive filter used in this study is of the form

$$H(z) = 1 - z^{-L_{\mathrm{I}}} F(z) \qquad (2)$$

where L_{I} is the integer part of the pitch period in samples and $F(z)$ is a fractional delay filter [15] that implements the non-integer part of the pitch period. We have used a second order Lagrange interpolator for the fractional delay filter. The use of the fractional delay filter is important for efficient cancelation of the voiced signal components. If only the integer part of the delay is used, the error is large especially with higher harmonics. This in turn will degrade the estimation of weaker voiced components.

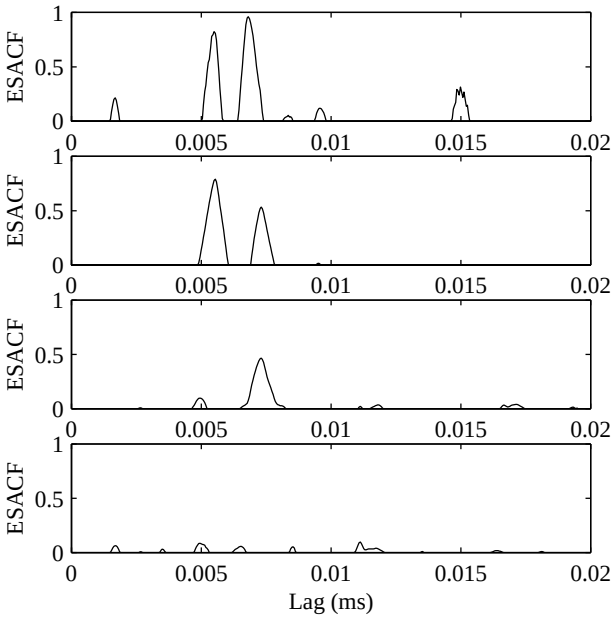

Figure 5: Example of iterative detection of voiced speech components. ESACF representations of the original signal with three components (top), after one component is extracted (second), after two components are extracted (third), and after three components are extracted (bottom).

Figure 5 shows an example of iterative analysis, detection, and extraction of voiced components that is applied to a mixture of three Finnish vowels `/a/` spoken by a male. The top plot depicts the ESACF representation computed on the original mixture. The most prominent component is identified, fundamental frequency estimate is fine-tuned using the NLS method, and the corresponding vowel is extracted for the second ESACF representation using the pitch predictive filter of Eq. 2. The iteration is performed twice more so that all the components are extracted in the bottom figure.

Notice that originally there were only two distinct peaks in the ESACF representation (cf. top plot). After the first component is removed, a peak corresponding to a weaker tone close to the removed peak appears in the second ESACF plot. This illustrates the potential accuracy gains available using iterative analysis and extraction.

4. SEPARATION EXAMPLES

The discrete pitch objects may be used to derive separate representations for harmonic sound sources. The pitch-lag locations provide estimates of the fundamental period, and the prominence values may be used to decide in which order the harmonic signals are segregated if an iterative algorithm is used.

In this study, we applied a comb-notch filtering method to separate two Finnish vowels /a/ and /i/ from a mixed voice signal. The length of the analyzed segment was 93 ms (2048 samples at 22 kHz sampling rate). The peak locations in the ESACF were used as fundamental period estimates, and the separated sound signals $s_{\mathrm{a}}(n)$ and $s_{\mathrm{i}}(n)$ were obtained by filtering the mixed signal with transfer functions $H_{\mathrm{i}}(z)$ and $H_{\mathrm{a}}(z)$, respectively. The two digital filters were designed to remove the pitch periodicities and they were implemented as $H(z) = (1 - z^{-P_{\mathrm{I}}} H_{\mathrm{f}}(z))^2$, where P_{I} is the integral part of the pitch period in samples and $H_{\mathrm{f}}(z)$ is a fractional delay filter [15] which implements the non-integral part of the pitch period. A second-order Lagrange filter

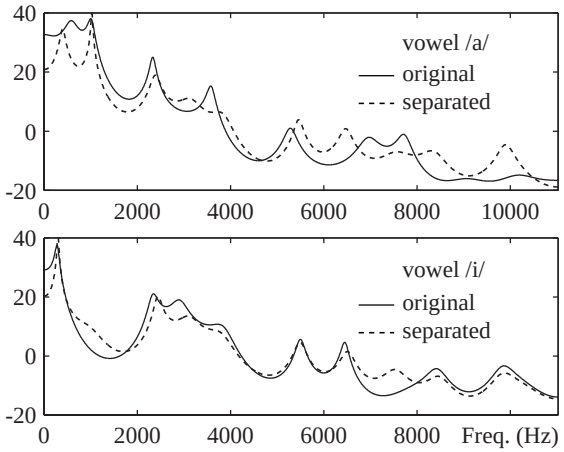

Figure 6: Separation of LP spectra from a mixture of two vowels. Top: original and reconstructed LP spectra of vowel /a/ with fundamental frequency of 129 Hz. Bottom: original and reconstructed LP spectra of vowel /i/ with fundamental frequency of 156 Hz.

was used for $H_{\mathrm{f}}(z)$.

Linear prediction (LP) spectra (order 24 at 22 kHz) of the separated signals $s_{\mathrm{a}}(n)$ and $s_{\mathrm{i}}(n)$ were computed to illustrate the separation results. For comparison, the LP spectra of the original vowels were also computed before they were mixed. Fig. 6 shows the results. The original (solid line) and the estimated (dashed line) LPs are depicted on the top for the /a/ sound and on the bottom for the /i/ sound. The example demonstrates that although a simple digital filter was used in the separation, the estimated LPs resemble the originals quite well. The performance may be improved by elaborating the design of separation filters more carefully.

5. DISCUSSION

In the examples shown above the method of multi-pitch analysis and separation works relatively well. Since complex nonlinear operations are used in SACF analysis, frame by frame, sometimes it misses a harmonic sound complex or generates spurious responses. This temporal variation, found also in clear steady-state cased, is worth further studies.

We have used a relatively long analysis frame and window. Shorter frames are found to be sensitive to noise and more severe spurious responses. It would be advantageous to increase the frame size but then the ability to follow fast pitch changes is compromised. There is need to develop methods that combine fast response and long temporal integration. Notice that the integration time constant of the human auditory system for many parameters is about 100–200 ms.

It seems evident that no straightforward methods for sound source separation that function properly for the general case exist. It is thus important to investigate principles that take into account the structural properties of sound or speech as perceived by the human auditory system at different levels. One such approach that looks promising in general is a multi-level predictive strategy as proposed in [11]. On the signal level we may apply linear prediction as discussed above. On parametric and higher structural levels we may do prediction as well but the representations are more object-based and non-numerical. The principles of auditory analysis [16], such as the old-plus-new strategy in segregating components, are needed.

In traditional speech processing the main emphasis has been on mathematical (often statistical) properties of the speech signal and the source. In many applications it appears important to attempt performance enhancement also by applying modeling of perception as well as structured and object-based representation of speech. This requires new methodologies that are yet to evolve.

6. ACKNOWLEDGMENT

This work has been financially supported by the GETA Graduate School at Helsinki University of Technology, the Foundation of Jenny and Antti Wihuri (Jenny ja Antti Wihurin rahasto), and Nokia Research Center.

7. REFERENCES

[1] W. M. Hartmann, "Pitch, periodicity, and auditory organization," *Journal of the Acoustical Society of America*, vol. 100, pp. 3491–3502, Dec. 1996.

[2] W. Hess, *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag, 1983.

[3] R. Meddis and L. O'Mard, "A unitary model for pitch perception," *Journal of the Acoustical Society of America*, vol. 102, pp. 1811–1820, Sept. 1997.

[4] R. Meddis and M. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: pitch identification," *Journal of the Acoustical Society of America*, vol. 89, pp. 2866–2882, June 1991.

[5] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, (Phoenix, Arizona), pp. 929–932, 1999.

[6] A. de Cheveignè and H. Kawahara, "Multiple period estimation and pitch perception model," *Speech Communication*, vol. 27, no. 3–4, pp. 175–185, 1999.

[7] R. Meddis and M. J. Hewitt, "Modeling of identification of concurrent vowels with different fundamental frequencies," *Journal of the Acoustical Society of America*, vol. 91, pp. 233–245, Jan. 1992.

[8] R. D. Patterson, "The sound of the sinusoid: Spectral models," *Journal of the Acoustical Society of America*, vol. 96, pp. 1409–1418, Sept 1994.

[9] U. K. Laine, M. Karjalainen, and T. Altosaar, "Warped linear prediction (WLP) in speech and audio processing," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. III.349 –III.352, 1994.

[10] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Heidelberg, Germany: Springer-Verlag, 1990.

[11] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, jun 1996.

[12] T. Tolonen, "Methods for separation of harmonic sound sources using sinusoidal modeling," in *AES 106th Convention*, (Munich, Germany), May 1999. To appear.

[13] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, New Jersey: Prentice Hall, 1997.

[14] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[15] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay—tools for fractional delay filter design," *IEEE Signal Processing Magazine*, vol. 13, pp. 30–60, Jan. 1996.

[16] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, Massachusetts, USA: The MIT Press, 1990.