SINGLE-CHANNEL SPEECH SEPARATION BASED ON MODULATION FREQUENCY

†*Lingyun Gu and* †‡*Richard M. Stern*

†Language Technologies Institute
‡Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, 15213, U.S.A.
{lgu, rms}@cs.cmu.edu

ABSTRACT

This paper describes an algorithm that performs a simple form of computational auditory scene analysis to separate multiple speech signals from one another on the basis of the modulation frequencies of the components. The most novel aspect of the algorithm is the use of the cross-correlation of the instantaneous frequencies of the components of a signal to identify and separate those components that are likely have been produced by a common sound source. The putative desired target speech signal is reconstructed by choosing those components that have the greatest mutual correlation, and then using extrinsic information such as fundamental frequency or speaker identification to determine which component clusters belong to which speaker. The system was evaluated by comparing speech recognition accuracy of a target speech signal that was extracted from a mixture of two speakers. It was found that recognition accuracy obtained when the separation was based on cross-correlation of changes in instantaneous frequency was better than the accuracy obtained when the separation was performed on the basis of fundamental frequency alone, for both the DARPA Resource Management Database and the Grid database used in the 2006 Speech Separation Challenge.[1]

Index Terms— speech analysis, speech recognition, time-frequency analysis, modulation frequency

I. INTRODUCTION

While the recognition of clean speech has been extensively examined and reasonably good recognition accuracy has been achieved, clean speech is often corrupted by different types of noise, such as background music, mechanical noise, or interfering speech. Mismatches between training and testing conditions produce degradation of recognition accuracy. Among the different types of additive noise, interfering speech is considered to be one of the most challenging sources of degradation, especially since speech-like interference is frequently difficult to distinguish from the target speech.

This paper describes a new approach to single-channel signal separation based on the detection of modulation frequencies, and the grouping of sets of frequencies that appear to be co-modulated, regardless of the extent to which they are harmonically related to each other.

[1]This research was supported by NSF Grant IIS-0420866.

I-A. Computational auditory scene analysis

Computational auditory scene analysis (CASA) seeks to exploit the inherent features contained in speech itself as cues to separate speech. Fundamental frequency, which is commonly referred to by its perceptual correlate pitch, is widely used in CASA systems as a cue to separate combined speech. As long as the pitch of the target and masking speech are different in a short time segment, the target speech can be resynthesized only from those components related to its pitch and harmonics. Multi-pitch tracking algorithm have been utilized to extract pairs of pitch trajectories (*e.g.* [1] [2]), and a CASA system based on pitch tracking was also built to separate speech based on estimated pitch (*e.g.* [3] [4]). One could develop a system to reconstruct the target speech from its fragments or perform the decoding based on the fragments alone (*e.g.* [5] [6]).

There are many different types of inherent information contained in speech, such as pitch, onset/offset, time/frequeny continuity, etc. Among all the various features, modulation frequency has been the object of much attention and has been used to detect pitch, separate speech, and modify speech, as will be discussed in the next section (*e.g.* [7]).

I-B. Uses of mdulation frequency

A narrowband signal can be represented by a higher frequency carriers modulated in amplitude and phase at lower frequencies. Consider, for example, the continuous-time representation of a sinusoid with time-varying amplitude $A(t)$ and phase $\theta(t)$,

$$y(t) = A(t)\cos(\theta(t)) = A(t)\cos(\omega_0 t + \int_0^t \omega(\tau)d\tau + \theta_0) \quad (1)$$

The instantaneous frequency $\omega_i(t)$ is the derivative of the instantaneous phase with respect to time

$$\omega_i(t) = \frac{d\theta}{dt} = \omega_0 + \omega(t) \quad (2)$$

where ω_0 is the carrier frequency and $\omega(t)$ is the modulation frequency, which represents deviations from the nominal frequency value ω_0. If a signal is a complex tone with multiple harmonics, the n^{th} harmonic of the fundamental would exhibit the same instantaneous frequency as the component at the fundamental frequency, but multiplied by the scaling factor n:

$$\omega_n(t) = n\omega_0 + n\omega(t) \quad (3)$$

Many studies have focused on modulation frequency in the fields of psychoacoustics and speech production, and it is beginning to be exploited for speech processing and recognition. For example,

the impact of modulation frequency on speech recognition was discussed in [8]. Filter design based on modulation frequency has also been proposed for preprocessing to mitigate the effects of reverberation [9]. Modulation frequency can also be combined with nonnegative matrix factorization to estimate pitch explicitly [10].

The potential exploitation of modulation frequency can contribute greatly to improved source separation. Traditional signal processing techniques such as short-time Fourier analysis (STFA) separate signals according to time and frequency. If modulation frequency can be consistently extracted, it provides a potential third orthogonal dimension along which signal components can be separated and clustered. The target speech signal can potentially be reconstructed by selecting those localized time-frequency regions that exhibit a particular common modulation frequency [11]. Nevertheless, accurate estimation of modulation frequency can be quite difficult for natural signals.

In the next section we will describe ways in which cross-channel correlation based on instantaneous frequency can be used to detect and cluster frequency bins that belong to a common sound source. Experimental results are discussed in Sec. III and our results are summarized in Sec. IV.

II. SIGNAL SEPARATION BASED ON MODULATION FREQUENCY

Voiced speech is perceived as having a pitch at the fundamental frequency of vibration of the vocal chords. Nevertheless, humans can never maintain truly constant pitch because of physiological limitations in the production mechanism, so the speech excitation signal is in practice non-stationary and only quasi-periodic.

While it is nominally worthwhile to estimate modulation frequency instead of (or in addition to) fundamental frequency, the modulation frequency of natural speech is limited to about 16 Hz, so changes evolve over a relatively long time, and one cannot observe a complete period of modulation without a very long analysis frame. Unfortunately, an analysis frame that is long enough to retain an entire modulation period will inevitably average out frequency fluctuations of shorter duration. Rather than attempting to estimate modulation frequency itself (as in [11]), we attempt to estimate correlations in the short-time frequency modulation over distinct frequency channels, with the goal of identifying frequency bins that are highly correlated with one another in terms of instantaneous frequency. This method requires the estimation of neither multi-pitch trajectories nor the modulation frequency itself. Instead, it provides system developers with the option of using either pitch or modulation as the basis for sound separation.

Figure 1 illustrates the system design. The combined speech $x[n]$ is first decomposed into a two-dimensional time-frequency representation $X_m[n,k]$ using short-time Fourier analysis (STFA), where m is the frame index, n is the time sample index within a frame and k is the index of frequency bins. Each frequency bin can be considered to be the output obtained by passing the original input through a narrow bandpass filter. The instantaneous amplitude and phase in each frequency bin will be slowly time varying. The phase information $\theta_m[n,k]$ is obtained easily from the inverse tangent of the quotient of the imaginary and real parts of the filter output $X_m[n,k]$:

$$\theta_m[n,k] = \arctan(\frac{imag(X_m[n,k])}{real(X_m[n,k])}) \tag{4}$$

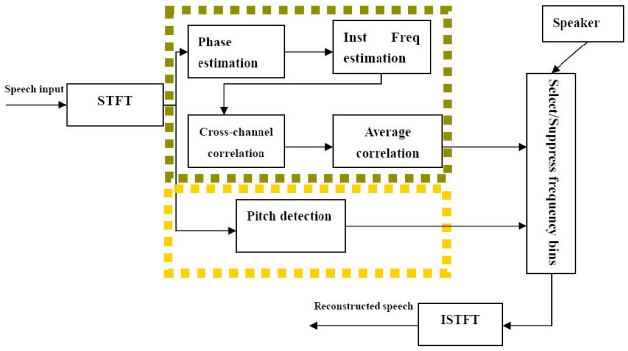

Fig. 1. System that implements the modulation-frequency-based (in the upper block) and pitch-based (in the lower block) speech separation algorithms.

The instantaneous frequency $\omega_{I,m}[n,k]$ is estimated by taking the first difference of the instantaneous phase, with care taken to deal appropriately with the effects of phase wrapping.

$$\omega_{I,m}[n,k] = \theta_m[n,k] - \theta_m[n-1,k] \tag{5}$$

Once the instantaneous frequency is obtained for each time-frequency location, a pair-wise short-time cross-channel correlation of instantaneous frequency is computed across all the frequency bins. $R_{\omega,m}(k_0,k_1)$ is the cross-channel correlation, evaluated for two frequency channels represented by the indices k_0 and k_1.

$$R_{\omega,m}(k_0,k_1) = \frac{C_m(k_0,k_1)}{\sqrt{C_m(k_0,k_0)C_m(k_1,k_1)}} \tag{6}$$

where $C_m(k_0,k_1)$ is the covariance of the instantaneous frequency in two bins k_0 and k_1:

$$\begin{aligned} C_m(k_0,k_1) = E[(\omega_{I,m}[n,k_0] - \overline{\omega_{I,m}[n,k_0]}) \\ (\omega_{I,m}[n,k_1] - \overline{\omega_{I,m}[n,k_1]}] \end{aligned} \tag{7}$$

where the statistical averages in the equations above are approximated by time averages within the analysis frame. This approach differs from that described in [1], in that we use cross-channel correlation over all frequency channels and do not use it to estimate fundamental frequency directly, while in [1] correlations are calculated only between adjacent frequency channels and correlation values are used as a tool to estimate pitch.

Figure 2 shows a cross-channel correlation pattern across all frequency bins for an actual voiced segment of speech with a fundamental of 135.5Hz. The small red-colored (bright) rectangles are frequency bins that are highly related to the fundamental frequency and all its harmonics. Since the correlation pattern is symmetric, either the horizontal or vertical axis can be used to collect the correlated frequency bins. If a frequency bin contains the fundamental frequency, it will have high correlation with all frequency bins containing its harmonics. Each column is averaged and the averaged value is compared to a preset threshold which is database dependent but not very sensitive. This threshold is always set to 0.25 in all of our experiments. Those frequency bins retained for further processing are selected from those elements of the cross-correlation representation for which the average correlation exceeds the preset value.

The cross-correlation matrix identifies components that belong together, but it does not indicate the speaker to which each frequency-component cluster belongs. A standard speaker identification (SID) system based on Gaussian mixture models (GMMs) [12] on a frame-by-frame basis to determine which frames are dominated by the target speaker or the interfering speaker. For each frame that is dominated by the target speaker, the frequency bins are selected that have high correlation with other frequencies, and the other frequencies are suppressed. For those frames that are identified as representing the interfering speaker, frequency bins with high correlation are suppressed while the other frames are retained. We note that no current algorithms deals perfectly with frames in which the fundamental frequency from different speakers is the same. Finally, a reconstruction of the target speaker is obtained through the inverse short-time Fourier transform of the retained time-frequency components only.

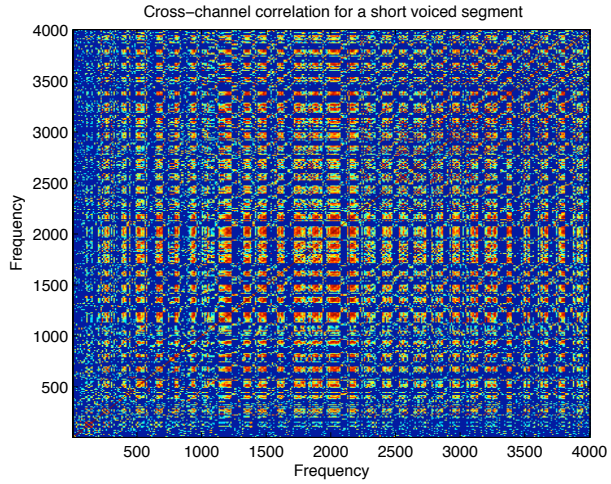

Fig. 2. Cross-channel correlation over all the frequency bins for a short voiced segment at frequency 135.5Hz, where red and yellow regions represent high correlation, while blue and green regions represent low correlation.

The system shown in Fig. 1 supports two parallel signal separation approaches, one based on correlations of modulation frequency (contained in the green box) and the other based on fundamental frequency alone (contined in the yellow box), both using the same peripheral signal processing and speech recognition software. The pitch-only separation algorithm uses de Cheveigné's YIN algorithm [2], which attempts to estimate the pitch value from the dominant speaker and resynthesize the entire utterance by only selecting the estimated pitch and its harmonics. In our experiments, the speech signal is windowed using 50-ms Hamming windows with 75% overlap for the modulation-frequency-based approach and 30-ms windows with 50% overlap for the pitch-only separation approach.

III. EXPERIMENTAL RESULTS

The modulation-frequency-based separation algorithm was evaluated by using the CMU SPHINX-III system and comparing its performance with three types of signal processing: (1) a baseline system using conventional MFCC processing, (2) the signal separation system based on modulation frequencies described in this paper, and (3) a simple signal separation system based on pitch tracking alone, for comparison. Two standard speech corpora were used for the evaluation, the Grid database which had been used used in the Speech Separation Challenge of 2006 [13] [14], and the familiar DARPA Resource Management (RM) database. The Grid corpus includes 34 speakers, each providing 500 clean utterances, for a total of 17,000 clean training utterances. The testing utterances are degraded at various SNRs. All utterances in the Grid database have a fixed format as specified in Table I. For example, a typical sentence could be "bin white at d 1 again". We only worked with speech with an SNR of $+6$ dB in the present work. The DARPA RM database consists of 1600 training utterances and 600 testing utterances. The vocabulary size of RM database is nominally 1000 words.

The evaluation methods for the two corpora are quite different. As noted above, sentences in the Grid database all consist of a verb followed by a color, a preposition, a letter, a digit, and an adverb in precisely that sequence. The only errors tabulated are substitutions of the digits and letters in the speech data. The task remains a difficult one, as digits and letters are frequently confused even by human listeners when competing speakers are present. We used the scoring tool prepared by the University of Sheffield to score recognition results for the Grid database.

Figure 3 summarizes the word recognition accuracies ($100\% - WER$) obtained using the Grid database, tabulated separately for interfering speakers of the same gender and the opposite gender of the target speaker. As is seen in the figure, separation based on modulation frequency provided better accuracy than separation based on pitch only and baseline MFCC processing. Differences between accuracies obtained using the MF and MFCC methods were always statistically significant, but the differences between scores obtained for the MF and Pitch methods are significant only for the 'different gender' and 'average' cases. The absolute level of performance was worse than the two CASA systems presented at Interspeech 2006 by groups at the University of Sheffield [6] and Ohio State University [4], but the present system does not yet include many of the modules that are known to be important in these more mature systems, including fragment decoding, sequential grouping, missing-feature reconstruction and methods for dealing with unvoiced segments.

VERB	COLOR	PREPOSITION	LETTER	DIGIT	ADVERB
bin	blue	at	a-z	1-9	again
lay	green	by	(no 'w')	and zero	now
place	red	on			please
set	white	with			soon

Table I. Structure of the sentences in the Grid database.

Since speaker ID information was not available for the RM database, we limited our comparisons to the 'different gender' case, using pitch tracking from the YIN algorithm only for gender separation in the MF configuration. Since same-gender competing speakers share a very similar pitch range with the target speaker, pitch estimation is not very useful for the same gender case. Conventional word error rates (WERs) were tabulated, including substitution, insertion, and deletion errors. Figure 4 summarizes the results for these experiments, and again separation based on modulation frequency provided the best recognition accuracy, with statistically significant differences between processing types. It is

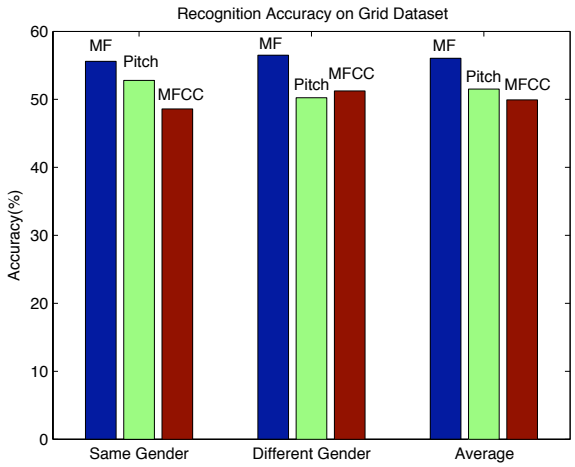

Fig. 3. Recognition accuracy obtained using the Grid database, comparing speech separation using modulation frequency (MF), speech separation using pitch only (Pitch), and baseline MFCC processing without speech separation (MFCC). The SNR in all three cases was 6 dB.

worth noting that both the modulation-frequency and pitch-only separation approaches provide worse performance than baseline MFCC processing for the two highest SNRs. The task of reducing WERs at low SNRs while retaining good performance for clean speech is still quite challenging.

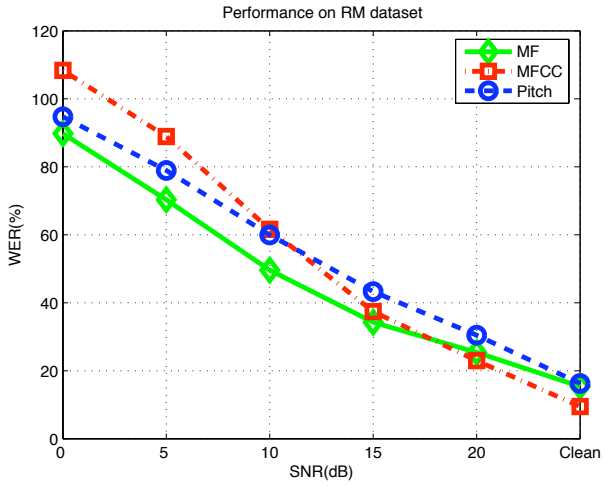

Fig. 4. Word error rates (WERs) obtained using the DARPA Resource Management Corpus, comparing speech separation using modulation frequency (MF), speech separation using pitch only (Pitch), and baseline MFCC processing without speech separation (MFCC).

IV. CONCLUSIONS

In this paper, a new single-channel speech-separation approach based on modulation-frequency detection and cross-channel correlation of instantaneous frequency is presented and evaluated. Using two different databases, we demonstrated that separation using instantaneous modulation frequency provided better recognition accuracy than separation based on fundamental frequency alone and baseline processing using MFCC features. We expect to observe further improvements by incorporating more accurate speaker identification and methods for dealing with unvoiced segments.

V. REFERENCES

[1] M. Wu, D.L. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 229–241, 2003.

[2] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of Acoustic Society of America*, vol. 111, pp. 1917–1930, 2002.

[3] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[4] S. Srinivasan, Y. Shao, Z. Jin, and D.L. Wang, "A computational auditory scene analysis system for robust speech recognition," in *Proceeding of Interspeech 2006*, 2006, pp. 73–76.

[5] J.P.Barker, M.P.Cooke, and D.P.W. Ellis, "Decoding speech in the presence of other sources," *speech communication*, vol. 45, pp. 5–25, 2005.

[6] J. Barker, A. Coy, N. Ma, and M. Cooke, "Recent advances in speech fragment decoding techniques," in *Proceeding of Interspeech 2006*, 2006, pp. 85–88.

[7] T. Chi, Y. Gao, M.C. Guyton, P.Ru, and Shihab Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, pp. 2719–2732, 1999.

[8] A. Takayuki, T. Mahoro, K. Noboru, T. Yukiko, and M. Yuji, "On the important modulation frequency bands of speech for human speaker recognition," in *Proceeding of Interspeech 2000*, 2000, pp. III774–III777.

[9] A. Kusumoto, T. Arai, T. Kitamura, M. Takahashi, and Y. Murahara, "Modulation enhancement of speech as a preprocessing for reverberantchambers with the hearing-impaired," in *Proceeding of ICASSP 2000*, 2000, pp. 853–856.

[10] F. Sha and L.K. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in *Proceeding of Advances in Neural Information Processing System 17*, 2005, pp. 1233–1240.

[11] S.M. Schimmel, L.E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Proceeding of ICASSP 2007*, 2007, pp. IV605–IV608.

[12] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identificiation using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[13] M. Cooke, "Grid corpus," Website, http://www.dcs.shef.ac.uk/spandh/gridcorpus/.

[14] M. Cooke, "Interspeech 2006 speech separation challenge," Website, 2006, http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm.