A HYBRID COHERENT-INCOHERENT METHOD OF MODULATION FILTERING FOR SINGLE CHANNEL SPEECH SEPARATION

A. Mahmoodzadeh[1], H. Sheikhzadeh[2], H. R. Abutalebi[1], H. Soltanian-Zadeh[3,4]

[1]Speech Proc. Research Lab, ECE Dept. Yazd University, Yazd, Iran
[2]EE Dept. Amirkabir University of Technology Tehran, Iran
[3]Control and Intelligent Processing Center of Excellence, University of Tehran, Tehran, Iran
[4]Image Analysis Lab. Henry Ford Health System, Detroit, USA
email: a.mahmoodzadeh@stu.yazduni.ac.ir

ABSTRACT

Single Channel Speech Separation has been the objective of extensive research in recent years. In this paper, we propose a hybrid system of coherent and incoherent modulation filtering for separation of the target speech from the interference. In the proposed system, subband envelopes are determined using a coherently detected subband carrier based on the time-dependent spectral Center-Of-Gravity (COG) demodulation and then the interference signal is eliminated by applying the adaptive Affine Projection (AP) filter to the subband envelop. The reference signal for the adaptive AP filter is provided by a parallel incoherent system of modulation filtering. Our evaluations, based on several objective measures, indicate that the proposed system extracts the majority of target speech signal segments with minimal interference, outperforming previous systems in voiced speech separation.

Index Terms— Affine Projection, Coherent modulation filtering, Instantaneous frequency, Speech separation.

1. INTRODUCTION

In real-world listening environments, speech reaching our ears is often corrupted by various types of acoustic interference. In recent years, single channel speech separation has been the object of extensive research for important applications such as telecommunication systems and Automatic Speech Recognition (ASR). The performance of these systems may severely degrade when speech is subjected to additive noise.

Major approaches to the single channel speech separation problem include: estimating the short-time spectra of interference [1], speech modeling, such as Gaussian Mixture Model(GMM) [2] or Hidden Markov Models [3], sparse decomposition [4], Non-negative Matrix Factorization (NMF) [5]. These methods usually assume certain properties of interference thus lacking the capacity for dealing with general acoustic interference.

Many natural and man-made signals can be represented as low frequency modulators which modulate higher frequency carriers. Therefore, modulation analysis and filtering is introduced as a tool for modifying, enhancing and separating narrowband analytic signals in recent years (e.g. [6]). Atlas and Janssen [7] present methods to separate audio from musical instruments with differing modulation frequency characteristics through modulation filtering. Also, in [8]-[9], the feasibility of target talker enhancement were demonstrated using modulation filtering. In these methods, by assuming that certain properties of interference are known, a fixed filter is used for speech enhancement. Obviously, there is no direct access to the interference signal in single channel scenarios.

Based on the above discussions, we propose a hybrid system based on the adaptive coherent modulation filtering for single channel speech separation. In the proposed system, first the subband signal is decomposed into modulator and carrier signals by using an estimate of the instantaneous frequency obtained by the time-dependent spectral Center-Of-Gravity (COG) introduced by Clark [10]. Modulation filtering is the process of modifying an analytic subband signal by filtering its modulator part and recombining the result with the original carrier. We employ an adaptive filter controlled by Affine Projection (AP) algorithm for each subband modulator signal to separate the target from the interference signal.

For the adaptive coherent modulation filtering, one has to estimate the desired signal which is not directly available in the single channel case. To solve this problem, we propose to employ an incoherent speech separation system to provide an approximation of the target speech signal. In each subband, the system estimates the pitch ranges of target and interference speakers using an onset-offset algorithm and then, for the speech separation objective, a soft mask is determined considering these pitch ranges and the modulation spectrum energy of the target and interference speakers in these ranges. To assess the quality of the separated signals, we report the separation results in terms of perceptual assessment of speech quality (PESQ),

weighted-slope spectral distance (WSS), log-likelihood ratio (LLR) and signal-to-noise ratio (SNR) as objective measures.

This paper is organized as follows. Section 2 describes the COG demodulation method. In Section 3, we provide a description of the proposed system. The results of the system on single channel speech separation are reported in Section 4. The paper concludes with a discussion in Section 5.

2. CENTER-OF-GRAVITY (COG) DEMODULATION

Time-varying spectral Center-Of-Gravity (COG) demodulates subband signals into complex modulator/carrier signals [10]. The signal product model of a discrete-time full-band signal $x[n]$ can be written as:

$$x[n]=\sum_{k=0}^{K-1} m_k[n]\cdot c_k[n], \tag{1}$$

where $m_k[n]$ and $c_k[n]$ are the modulator and carrier signals of the k^{th} subband from a filterbank with K channels. Demodulation is defined as the process of estimating $m_k[n]$ and $c_k[n]$ for the given $x[n]$ for all n and k. The complex carrier is defined as:

$$c_k[n]=\exp(j\phi_k[n]), \tag{2}$$

where $\phi_k[n]$ is the phase of the carrier $c_k[n]$. Hence, the complex modulator via the demodulation rule is obtained as:

$$m_k[n]=x_k[n]\cdot c_k^*[n], \tag{3}$$

where $x_k[n]$ is the k^{th} analytic band-pass signal. To find the modulator, we need only to define the phase of the carrier, $\phi_k[n]$. Typically, the so-called instantaneous frequency of an oscillating signal $c_k[n]$ is defined as the derivative of the phase. In discrete-time domain, we approximate the derivative with a first-order difference,

$$f_k[n]=\phi_k[n]-\phi_k[n-1], \qquad \phi_k[-1]=0 \tag{4}$$

where the carrier frequency $f_k[n]$ is in radians/sample. The first-order difference is a rather crude approximation to the derivative, but it has the advantage of being invertible via the cumulative sum,

$$\phi_k[n]=\sum_{p=0}^{n} f_k[p] \tag{5}$$

Therefore, determination of the instantaneous frequency of the subband is necessary for calculating the modulator and carrier signals. The COG approach defines $f_k[n]$ as the average frequency of the instantaneous spectrum of $x_k[n]$ at time n. Conceptually we estimate the instantaneous spectrum using the Short-Time Fourier Transform (STFT),

$$X_k[i,n]=\sum_{p} g[p]x_k[n+p]e^{-j2\pi(i/L)p}, \quad i=0,\ldots,L-1 \tag{6}$$

where $g[p]$ is a short spectral-estimation window and L is the STFT size. Then, the center-of-gravity is defined as:

$$f_k[n]=\frac{\sum_{i=0}^{L-1} r[i]\left|X_k[i,n]\right|^2}{\sum_{i=0}^{L-1}\left|X_k[i,n]\right|^2}, \tag{7}$$

where $r[i]$ is a weighting function and defined as:

$$r[i]=\begin{cases} 2\pi i/L, & 0\le i\le L/2 \\ 2\pi i/L-2\pi, & L/2<i<L \end{cases} \tag{8}$$

3. PROPOSED SYSTEM DESCRIPTION

The main idea behind our system is the segregation of voiced target speech signal from the interference based on adaptive coherent modulation filtering in single channel. The overall multi-stage proposed system is depicted in Fig. 1. The signal processing of the adaptive coherent modulation filtering relies on the determining of the modulator and carrier signals of the mixture subband signal. Therefore for coherent demodulation using the COG method, at first, the wideband mixture signal is decomposed into narrowband subbands using a STFT filterbank (see Fig. 1). After applying the filterbank on the mixture signal, each subband signal is decomposed into modulator and carrier signals using the COG coherent demodulation method. Then, for separating the target modulator signal from the interference signal, the modulator signal of each subband is filtered using the adaptive AP filter.

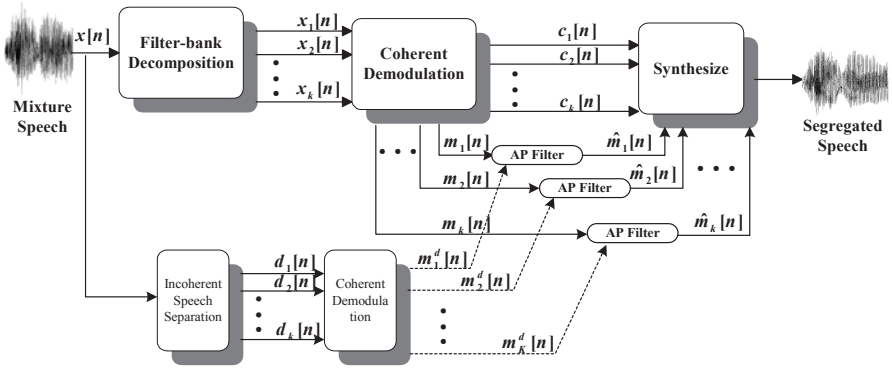

Fig. 1. Block diagram of the proposed system.

In each subband, the filter coefficients of the adaptive filter are determined by the subband Affine Projection algorithm [11] minimizing the squared error, E_k , between the filtered modulator signal, $\hat{m}_k[n]$, and the desired modulator signal, $m_k^d[n]$:

$$E_k = \sum_n \left| \hat{m}_k[n] - m_k^d[n] \right|^2 \quad (9)$$

Since the input signals to the subband adaptive filters are narrowband, we employ the AP algorithm which is a very fast and efficient adaptive algorithm especially when the input signal is colored. As we know, the reference channel is necessary for adaptive AP filtering; however, this channel is not available in the single channel case. To solve this problem, the separated target signal is used as the reference channel obtained from the incoherent speech separation system proposed in [12]-[14], as shown in Fig. 1. The desired modulator signal is obtained by applying the COG method on the separated target signal (reference channel).

Generally, two types of envelop detectors can be distinguished: coherent and incoherent detectors. The incoherent detectors estimate the modulators using a magnitude or magnitude-like operator, whereas the coherent detectors employ a carrier estimator based on an instantaneous frequency property of the signal.

The incoherent envelope detectors have the serious limitation that the subband modulator and carrier signal generally exceed the bandwidth of the subband signal and therefore, the signal's modulator domain is not closed under convolution. On the other hand, coherent detectors are able to adjust the bandwidth of the carrier by estimating the instantaneous frequency of the signal, and thereby adjust the bandwidth of the modulator. In addition, the complex modulator signal is closed under convolution.

In spite of these limitations, the incoherent modulation transform is able to create a modulation spectrum with wide frequency bandwidth in modulation frequency domain. For the speech signal, this property is used to resolve the pitch frequency range of two co-channel talkers in the modulation frequency domain, since this range is largely non-overlapping in this domain [8].

Fig. 2 depicts the general schematics of the employed multi-stage incoherent system (already proposed by the authors in [12,14]).The incoherent speech separation firstly estimates the pitch range of target speech in each frame of data in the modulation frequency domain by using an onset-offset algorithm and then uses the estimated pitch range to segregate target speech. Speech separation is performed by filtering the mixture signal with a mask extracted from the modulation spectrogram.

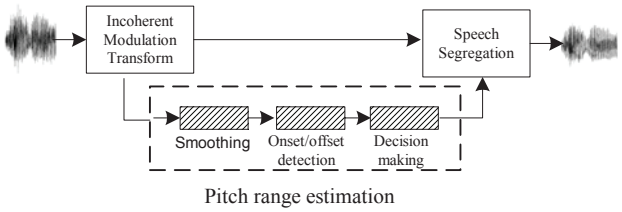

Fig. 2. Block diagram of the incoherent speech separation system.

4. EXPERIMENTAL RESULTS

As a proof of concept, we evaluate the separation performance of the proposed system with the database of a speech corpus and interference [15]. The database contains utterances from both male and female speakers. We mix these utterances with interferences (filtered through the room impulse response) at different SNR levels. The interference signals are: N0) 1 kHz pure tone, N1) white noise, N2) noise bursts, N3) babble noise, N4) rock music, N5) siren, N6) trill telephone, N7) female speech, N8) male speech and N9) female speech.

The input signal is digitized at a 16 kHz sampling rate. The filterbank has 256 filter channels with a prototype Hanning filter of 32 ms long and a frame rate of 8 ms. The separation results are quantified using PESQ (ITUT standard P.862.1), WSS and LLR as the objective speech quality measurement [16]. Also, the separation SNR of the proposed method is compared with other state-of-the-art methods: Hu and Wang system [17] (as a feature based method) and the spectral subtraction method [18].

Tables 1 and 2, present the performance of the proposed speech separation system in terms of objective measures: PESQ, WSS, and LLR for different SNR (dB) situations. The results shown are averaged for separated signal from the mixture of a target male speaker with a) a babble noise and b) a male speaker. Each table includes two cases; case 1: the proposed system when the clean speech signal is actually provided for the AP filter, and case 2: the proposed system when employing the separated speech obtained by the incoherent speech separation system for the AP filter. By comparing the results of cases 1 and 2 in the tables, one may conclude that the performance of the proposed coherent separation system is satisfactory; particularly, when speech is the interference signal.

The average SNR for each intrusion is shown for the proposed system in Fig. 3, compared with those of the original mixtures, Hu and Wang's system [17], and a spectral subtraction method [18]. The proposed system consistently outperforms the Hu and Wang system as well as the spectral subtraction. On average, the proposed system obtains a 17.05 dB SNR gain, which is about 2.12 dB better than Hu and Wang system and 8.6 dB better than spectral subtraction.

Table 1. Speech separation results versus different SNR's for a mixture of male target speaker and babble noise in terms of objective measures LLR, WSS, and PESQ.

SNR(dB)		-10	-5	0	5	10	15
Case 1	PESQ	2.39	2.80	3.23	3.50	3.65	3.79
	WSS	40.74	36.19	29.34	24.64	20.54	17.79
	LLR	0.41	0.49	0.50	0.52	0.51	0.50
Case 2	PESQ	2.32	2.89	3.08	3.44	3.50	3.37
	WSS	69.68	60.25	53.17	41.89	37.22	34.54
	LLR	0.76	0.66	0.66	0.70	0.70	0.64

Table 2. Speech separation results versus different SNR's for a mixture of male target and interference speakers in terms of objective measures LLR, WSS, and PESQ.

SNR(dB)		-10	-5	0	5	10	15
Case 1	PESQ	2.44	3.00	3.37	3.52	3.68	3.85
	WSS	41.72	26.08	37.26	33.80	26.66	16.75
	LLR	0.37	0.46	0.44	0.42	0.47	0.54
Case 2	PESQ	2.60	2.64	3.21	3.36	3.40	3.43
	WSS	52.72	48.64	44.34	32.25	31.15	29.93
	LLR	0.63	0.62	0.63	0.66	0.76	0.61

5. CONCLUSIONS

In this paper, we presented a new hybrid coherent modulation filtering system employing the adaptive affine projection for filtering the modulator signal of each subband signal to separate the target signal from the interference. Also we employed the separated signal obtained from the incoherent speech separation as an estimate of clean speech for adaptive filtering. Employing the objective speech quality measurement, it was observed that the perceived speech quality of the proposed system was acceptable. Also, through several experiments it was observed that the proposed method was superior to the Hu and Wang and the spectral subtraction separation systems.

REFERENCES

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Process.*, vol. 9, pp. 504-512, 2001.

[2] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, 2007.

[3] L. Benaroya, and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," *Proc.* Independent Component Analysis, pp. 957–961, Japan, April 2003.

[4] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," *Proc. int. Comput. Music Conf.*, pp. 231-234, 2003.

[5] M. Hele´n, and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," *Proceedings of the European Signal Processing Conference*, Turkey, Sept. 2005.

[6] A. Kusumoto, T. Arai, T. Kitamura, M. Takahasi, and Y. Murahara, "Modulation enhancement of speech as preprocessing for reverberant chambers with the hearing-impaired," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 853–856, 2000.

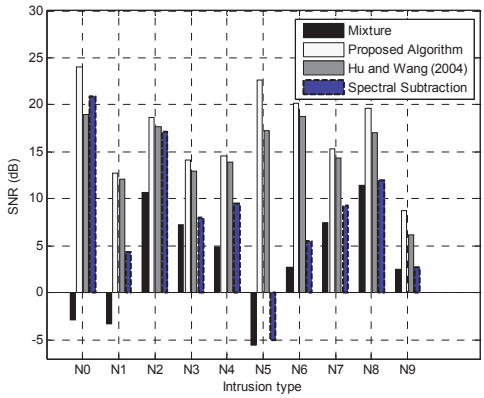

Fig. 3. SNR results for segregated speech and original mixtures for a corpus of voiced speech and various intrusions.

[7] L. Atlas, and C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 461-464, 2005.

[8] S.M. Schimmel, L.E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 605-608, 2007.

[9] Q. Li, and L. Atlas, "Coherent modulation filtering for speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 4481-4484, 2008.

[10] P. Clark, and L. Atlas, "Time-frequency coherent modulation filtering of non-stationary signals," *IEEE Trans. Sig. Process.*, vol. 57, no. 11, pp. 4323-4332, Nov. 2009.

[11] H.R. Abutalebi, H. Sheikhzadeh, R.L. Brennan, and G.H. Freeman, "Affine projection algorithm for oversampled subband adaptive filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 209-212, 2003.

[12] A. Mahmoodzadeh, H.R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh, "Single Channel Speech Separation with a Frame-based Pitch Range Estimation Method in Modulation Frequency," *IEEE International Symposium on Telecommunications (IST),* pp. 609-613, 2010.

[13] A. Mahmoodzadeh, H.R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh, "Determination of Pitch Range Based on Onset and Offset Analysis in Modulation Frequency Domain," *IEEE International Symposium on Telecommunications (IST)*, pp. 604-608 ,2010.

[14] A. Mahmoodzadeh, H.R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh, "Single channel speech separation in modulation frequency domain based on a novel pitch range estimation method," submitted to *EURASIP Journal on Applied Signal Processing*, under revision.

[15] M.P. Cooke, *Modeling auditory processing and organization. Cambridge*, U.K: Cambridge Univ. Press, 1993.

[16] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Rec*. p. 862, 2001.

[17] G. Hu, and D.L. Wang, "Monaural speech separation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Net.*, vol. 15, pp. 1135-1150, 2004.

[18] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: a guide to theory, algorithms, and system development*, Upper Saddle River NJ: Prentice Hall PTR, 2001.