Single Channel Speaker Segregation using Sinusoidal Residual Modeling

Rajesh M Hegde and A. Srinivas
Dept. of Electrical Engineering
Indian Institute of Technology Kanpur
Email: rhegde@iitk.ac.in

***Abstract*—In this paper we address the two speaker segregation problem in a single channel paradigm using sinusoidal residual modeling. An appropriate selection of the number of sine waves, window length and hysteresis threshold, is done so as to model and synthesize the underlying signal corresponding to the speaker with the lower pitch period, using an amplitude only sine wave synthesis. The sinusoidal residual is then computed after restimating the phases with known amplitudes, by minimizing a criterion function. This residual corresponds to the the speaker with the higher pitch period. But such a residual consists of harmonic components of the speaker with the lower pitch period. We therefore estimate a binary mask from the spectrograms of the synthesized signal and the residual using a min-max technique to further improve the quality of the segregated speech. This segregation technique is then integrated into a co-channel speaker identification system, at various target to interference ratios. Reasonable improvements in identification performance are noted from these experiments.**

I. Introduction

Recovering individual speech signals from a combination of two or more sources is a becoming a central problem in speech processing. Several approaches [1], have been tried to solve this problem ranging from the use of spatial information [2], to incorporating visual information [3], along with speech. But single channel speech segregation without the prior knowledge of the speech sources is challenging. In this paper we attempt to segregate the individual speakers from a mixture of two speakers collected over a single microphone. In Section II and III, we describe the sinusoidal residual modeling technique [4], and a formulation of the two speaker segregation problem respectively. The additive bank of sine wave synthesis of the estimated amplitudes and frequencies using this model results in a signal that corresponds to the source with the lower pitch period. But it contains some background information corresponding to the second speaker with the higher pitch period. The sinusoidal residual computed using a synthesis after restimating the phases results in the source with the higher pitch period. We attempt to illustrate and justify the reasons for the sinusoidal residual to contain information of the source with a higher pitch period in Section III-A. From the synthesized signal and the residual, we derive a mask using the min-max technique. This mask is applied on the synthesized signal to further refine the quality of the segregated sources. The computation of the mask and subsequent results are discussed in Section IV-A. This method is then integrated into a co-channel speaker identification system in Section V. The limitations of the technique and conclusions are discussed in in Section VI.

II. Sinusoidal Modeling

Sinusoidal modeling is based on the model suggested by Quatieri and McAulay [4], where a speech signal can be represented by a sum of amplitude-frequency modulated sine waves. The speech signal $x(n)$ can be expressed as a sum of time varying frequencies, amplitudes and phases as

$$x(n) = \sum_{k=0}^{N-1} a_k(n) cos(2\pi n f_k(n) + \theta_k(n)) \tag{1}$$

where $a_k(n)$, $f_k(n)$, and $\theta_k(n)$ are the amplitudes, frequencies and the phases of the speech signal which are all slowly varying functions of n. The underlying number of sine waves that can be used to reasonably represent the speech signal is given by N. The short time Fourier transform (STFT) of $x(n)$ after applying a window $w(n)$ on the speech signal $x(n)$ [5], is given by

$$X[k,n_0] = \sum_{n=0}^{N-1} x(n+n_0)w(n)e^{-j(\frac{2\pi nk}{N})} \tag{2}$$

The STFT $X[k,n_0)$ can also be expressed as

$$X[k,n_0] = |X\left(\frac{2\pi nk}{n_0}\right)|e^{j\theta\left(\frac{2\pi nk}{n_0}\right)} \tag{3}$$

In Equation 2, n_0 is the hop size at which the Fourier transform is evaluated. In Equation 3, $|X\left(\frac{2\pi nk}{n_0}\right)|$ corresponds to the short time magnitude spectrum and $\theta\left(\frac{2\pi nk}{n_0}\right)$ corresponds to the phase spectrum.

III. Formulation of the two speaker segregation problem using sinusoidal modeling

A mixture of two speech signals can be represented by the sum of two sets of sinusoids. Each set consists of multiple sinusoids, each with time varying amplitudes, frequencies, and phases [4]. Let x(n) represent the mixture of two speakers represented by $x_a(n)$ and $x_b(n)$ such that

$$x(n) = x_a(n) + x_b(n) \tag{4}$$

where

$$x_a(n) = \sum_{k=1}^{K_a} a_k(n) cos(\omega_{a,k}n + \theta_{a,k}) \tag{5}$$

and

$$x_b(n) = \sum_{k=1}^{K_b} b_k(n)cos(\omega_{b,k}n + \theta_{b,k}) \quad (6)$$

where $a_k, \omega_{a,k}, \theta_{a,k}$ are the amplitudes, frequencies, and phases respectively of the first speaker. A similar parameter set can be associated with the second speaker as in Equation 6. The windowed speech mixture can therefore be represented by

$$x_w(n) = w(n)[x_a(n) + x_b(n)] \quad (7)$$

where w(n), is the non zero analysis window which is centered and symmetric about the time origin. The Fourier transform of the windowed mixture $x_w(n)$ is given by

$$X_w(\omega) = \sum_{k=1}^{K_a} a_k(n)e^{j\theta_{a,k}}W(\omega-\omega_{a,k}) + b_k(n)e^{j\theta_{b,k}}W(\omega-\omega_{b,k}) \quad (8)$$

by making a reasonable assumption that negative frequency contribution is negligible and scale factors if any have been absorbed by the window term in Equation 8. Using Equations 4 through 8, it is conceivable that the two voice waveform can be reconstructed by selecting the appropriate number of sine waves and window length for the underlying pitch of each speaker. In addition, the two speaker case has the additional constraint that the analysis window length be chosen so as to resolve frequencies more closely spaced when compared to the one speaker case.

A. Significance of frame size, frame rate, hysteresis and pitch

Generally, for modeling a single speaker, the window size $w(n)$, as in Equation 7, is selected to get the the adequate resolution in the time frequency representation such that all harmonics are resolved. If it is too long there is a loss in time resolution as this limits the amplitude rate of change of the sinusoids. The hop size is selected as half the window size. It is worthwhile to note here that sinusoidal modeling of a single speaker uses window lengths which are 2 to 3 times the lowest fundamental period (pitch). Since we are primarily interested in modeling one of the two speakers, we use window lengths that accurately model the speech source with a lower pitch period. The sources considered here exhibit pitch in the range of 60 HZ (male) to 120 Hz (female). The hysteresis [4], threshold is also adjusted such that the peak tracking mechanism described in [4], picks peaks corresponding to the source with a lower pitch period and the corresponding speech signal synthesized using an amplitude only synthesis. This speech waveform estimate of the speaker with the lower pitch period is then subtracted from the original mixture to compute the waveform estimate of the speaker with the higher pitch period. The perceptual quality of the resulting residual is not adequate and therefore we follow a method of restimating phases in Section III-B.

B. Computing the sinusoidal residual by restimating phases

The mixture of two speakers is first modeled using sinusoidal analysis [4], by selecting the appropriate frame size, hop size and the hysteresis thresholds such that the speaker with a lower pitch period is modeled. This analysis results in a set of sinusoidal tracks which define contours in amplitudes, phases, and frequencies. The amplitudes and frequencies are used to synthesize the signal using amplitude only sine wave synthesis [6]. The signal synthesized using such a amplitude only (non phase preserving) synthesis, in each time frame of length N, is given by

$$x_{sa}(n) = \sum_{k=0}^{N-1} a_k(n)cos(2\pi f_k(n)) \quad (9)$$

This signal models the speaker with a lower pitch period to reasonably well but retains some background information corresponding to the other speaker. Hence we re synthesize this signal using both amplitudes and phases as

$$x_{sp}(n) = \sum_{k=0}^{N-1} a_k(n)cos(2\pi n f_k(n) + \theta_k(n)) \quad (10)$$

where $\theta_k(n)$ are the restimated phases computed on a frame wise basis by minimizing the error criterion

$$\min_{\theta_k} ||x(n) - x_{sp}(n)|| \quad (11)$$

The above problem can be written in matrix form as

$$X = H\Theta \quad (12)$$

where $X = [x(0), x(1), x(2), ..., x(N-1)]^T$,
$\Theta = [\theta(0), \theta(1), .., \theta(N-1)]^T$, and H is given by

$$H = \begin{pmatrix} h_0(1) & h_1(1) & . \; . & h_{N-1}(1) \\ h_0(2) & h_1(2) & . \; . & h_{N-1}(2) \\ . & . & . \; . & . \\ . & . & . \; . & . \\ . & . & . \; . & . \\ h_0(N-1) & h_1(N-1) & . \; . & h_{N-1}(N-1) \end{pmatrix}$$

The elements of the matrix H are computed as

$$h_k(n) = \sum_{n=0}^{M-1} a_k e^{j\omega_k n} \quad (13)$$

where $n = 0, 1, ..., (M-1)$. Since A is a full matrix and $H^H H$ is invertible, the solution to Θ in the least squares sense is given by

$$\Theta = ((H^H H)^{-1} H^H X \quad (14)$$

where matrices X, H, and Θ are as defined earlier. The values of Θ thus estimated are used to synthesize the primary sinusoidal signal (estimate of the speaker with a lower pitch period) as in Equation 10. The sinusoidal residual residual for each time frame after restimating the phases is now computed as

$$e_s(n) = [x(n) - x_{sp}(n] \quad (15)$$

where $e_s(n)$ is the estimate of the speaker with a higher pitch period.

C. Modeling a mixture of two speakers using sinusoidal residual modeling

We consider a mixture of two speakers to illustrate the approach described in Section III-B. For ease of comparison, the mixture is selected exactly as in [7], and consists of a male and a female speaker. The results of sinusoidal residual modeling of the two sources are illustrated in Figure 1. The waveform of the mixture considered is shown in Figure 1 (a), and its spectrogram is shown in Figure 1 (b). The sinusoidal tracks or the partials computed from the mixture and the corresponding residual after phase reestimation are shown in Figure 1 (c) and (d) respectively. It can be noted from the Figure 1 (c) and (d) that the sinusoidal tracks of the mixture and the residual model the individual sources present in the mixture. To further clarify these results we compute the spectrograms of the signals synthesized using the tracks shown in Figure 1 (c) and also the residual. The original waveform of the mixture, its corresponding spectrogram, the spectrogram of the signal synthesized using the tracks shown in Figure 1 (c), and the spectrogram of the residual computed after the phase restimation process are shown in Figure 2 (a), (b), (c), and (d) respectively.

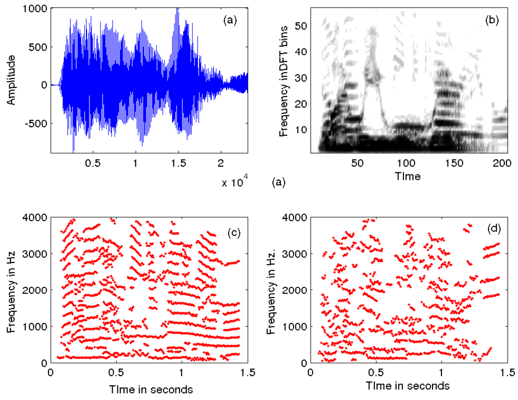

Fig. 1. Illustration of the sinusoidal tracks for the mixture and the corresponding residual. (a) Example Waveform of the mixture of two speech sources, (b) Spectrogram of the waveform in (a), (c) Sinusoidal tracks of the mixture of two speech sources in (a), and (d) Sinusoidal tracks of the residual.

IV. Speaker separation using sinusoidal residual modeling

Two sentences from the GRID corpus [8] are considered. The first sentence is from a female speaker uttering the sentence */Place red in a zero now/*. The second sentence is from a male speaker uttering */Set white with p two soon/*. The two sentences are added to generate a mixture of the two sources. Sinusoidal residual analysis and synthesis is employed to synthesize two signals, namely the amplitude only synthesized signal and the residual after phase restimation. The results are illustrated in Figures 3 and 4. In Figure 3 (a), is shown the spectrogram of the mixture of the two speakers. The original spectrogram of the female speaker is shown in Figure 3 (b), while the spectrogram of the signal synthesized using amplitude only sinusoidal synthesis is shown in Figure 3 (c). The original spectrogram of the male speaker is shown in Figure 3 (d), while the spectrogram of the residual computed after phase restimation signal is shown in Figure 3 (e). It is significant to note that the spectrogram of the signal synthesized using amplitude only sinusoidal synthesis is similar to the female speakers original spectrogram while the spectrogram of the sinusoidal residual is similar to the spectrogram of the male speaker. But on hearing the individual signals it is noticed that the amplitude only sinusoidal synthesized signal has some background information corresponding to the female speaker. To refine this signal further we compute a mask using min-max technique.

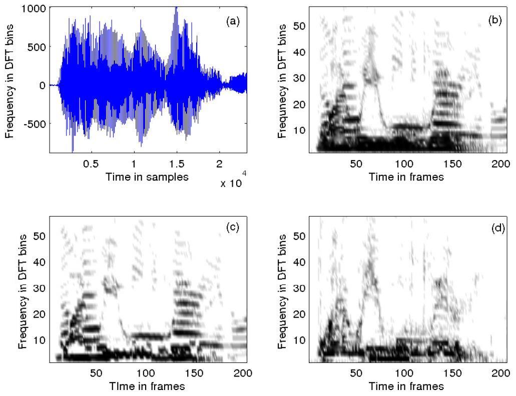

Fig. 2. Spectrograms of the mixture, the sinusoidal model and the corresponding residual. (a) Example Waveform of the mixture of two speech sources, (b) Spectrogram of the waveform in (a), (c) Spectrogram of the signal synthesized from the sinusoidal tracks, and (d) Spectrogram of the residual after phase restimation.

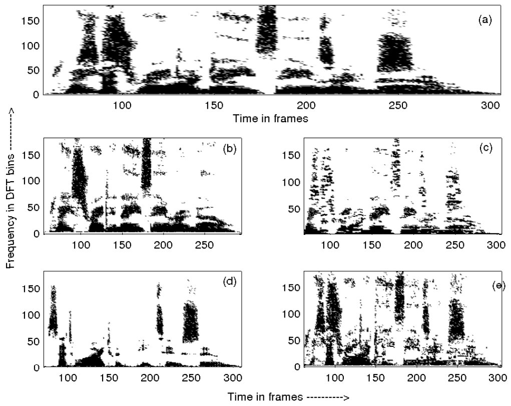

Fig. 3. Spectrograms of the mixture, the sinusoidal model and the corresponding residual. (a) Spectrogram of the mixture of two speech sources, (b) Spectrogram of the original female speaker, (c) Spectrogram of the signal synthesized using amplitude only synthesis, (c) Spectrogram of the original male speaker, and (d) Spectrogram of the residual computed after the phase restimation process.

A. Mask estimation

It should be noted here that only speech sources whose pitch tracks are reasonably separated are selected. As noted in Section IV, the signal synthesized using amplitude only synthesis does retain the information of the source with the higher pitch period. To further improve the quality of this signal we estimate a spectrographic mask using min-max method as follows

$$\begin{aligned} mask(\omega,t) &= 1; if X_A(\omega,t) > X_r(\omega,t) \\ mask(\omega,t) &= 0; if X_A(\omega,t) < X_r(\omega,t) \end{aligned} \qquad (16)$$

where $mask(\omega,t)$ is a spectrographic mask as a function of frequency and time, $X_A(\omega,t)$, is the log magnitude of the spectrogram of the amplitude only sinusoidal synthesized signal, and $X_r(\omega,t)$ is the spectrogram of the residual computed after phase restimation. Note that in Equation 16, $\omega = \frac{2\pi nk}{N}$, and, N is the short window over which the mask is estimated. The mask thus estimated is applied to the signal synthesized using amplitude only sinusoidal synthesis which results in a improved version of the original signal corresponding to the female speaker. The spectrograms of the original signal corresponding to the female speaker and the refined signal after applying the mask are shown in Figure 4 and 5 respectively.

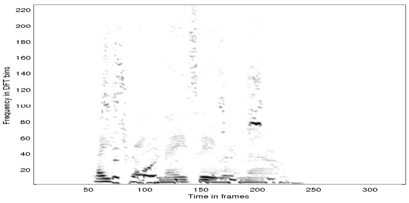

Fig. 4. Spectrogram of the amplitude only sinusoidally synthesized signal corresponding to the female speaker.

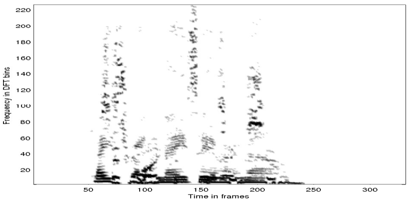

Fig. 5. Spectrogram of the refined signal corresponding to the female speaker.

V. Experiments on co-channel speaker identification

The speaker segregation technique is applied to the task of co-channel speaker identification [9], using data from the TIMIT database [10]. We consider 50 speakers from the same ('DR1') dialect region consisting of 20 male and 30 female speakers. For building each speaker model (a GMM with 128 mixtures), 8 sentences from that particular speaker are used. The other 2 sentences are used for testing. For simulating co-channel speaker mixtures, these two sentences from one speaker (target speaker) are mixed with the other (interfering speaker) at various target to interference ratios (TIR). The different target to interference ratios are obtained by scaling the speech of the interfering speaker. The TIR is calculated as

$$\mathbf{TIR} = \frac{\sigma_{\mathbf{T}}^{\mathbf{2}}}{\sigma_{\mathbf{I}}^{\mathbf{2}}} \qquad (17)$$

where σ_T^2 and σ_I^2 are the variances of the target and the interfering speakers respectively, and are computed across all frames of the test utterance. Table I, lists the results of identification of either the target or the interfering speaker from a mixture of 2 speakers using direct (no separation), using sinusoidal residual modeling for two speakers, as proposed in Section IV, and also using the sinusoidal residual modeling followed by masking, as in Section IV-A. The results are listed as % speaker error rate (SER). The target speaker identification error rate (TSER) is also listed in Table II. TSER corresponds to error rate for identifying the target speaker only. It is noted that the technique of sinusoidal residual modeling followed by masking gives a reasonable improvement in recognition performance when compared to direct and simple sinusoidal modeling. Secondly the improvements are noticed across all TIRs.

TABLE I
Experimental results of co-channel speaker identification as Target/Interfering SER

Technique	TIR	% Target/Interfering SER
Direct	-5dB	64
	0 dB	52
	5 dB	45
	10 dB	36
SR Modeling	-5 dB	7
	0 dB	6
	5 dB	5
	10 dB	2
SR Modeling + Masking	-5 dB	6
	0 dB	5
	5 dB	4
	10 dB	1

TABLE II
Experimental results of co-channel speaker identification as Target SER

Technique	TIR	% Target SER
Direct	-5dB	82
	0 dB	78
	5 dB	72
	10 dB	68
SR Modeling	-5 dB	48
	0 dB	35
	5 dB	24
	10 dB	18
SR Modeling + Masking	-5 dB	46
	0 dB	32
	5 dB	20
	10 dB	16

VI. Conclusion

New methods of blind single channel speaker segregation based on the sinusoidal residual modeling are discussed in this paper. Masks estimated using the amplitude only synthesized signal and the residual, further improve the separation. The initial results of this technique applied to co-channel speaker identification are promising. We are currently addressing methods of efficient multipitch estimation, selection of window and hop size based on multipitch estimation, efficient mask estimation and other related issues to improve the quality of the separated signals. Segregation of undesired stationary and non stationary events from the speech signal using this technique and the use such a separated speech signal in a conventional speech recognition system is another issue that is currently being addressed by us.

References

[1] A. J. Bell, and T. J. Sejnowski, "An information- maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[2] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, US, 2001.

[3] J. Hershey, and M. Casey, "Audio-visual sound separation via hidden markov models," *Neural Information Processing Systems*, 2001.

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

[5] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993.

[6] E. R. Remez, P. E. Rubin, D. B. Pisoni,and T. D. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, pp. 947–950, 1981.

[7] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech, Audio Processing*, vol. 7, pp. 229–241, 2003.

[8] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *available at: http://www.dcs.shef.ac.uk/spandh/gridcorpus/.*

[9] D. P. Morgan, E. B. George, L.T Lee and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2003, vol. 2, pp. 205–208.

[10] NTIS, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.