AN ALGORITHM FOR SPEECH SEGREGATION OF CO-CHANNEL SPEECH

Srikanth Vishnubhotla, Carol Y Espy-Wilson
(srikanth@umd.edu , espy@umd.edu)
Institute for Systems Research & Department of Electrical & Computer Engineering,
University of Maryland, College Park, MD, USA

ABSTRACT

This paper introduces an algorithm to separate speech streams from a single-channel speech mixture. Most current speech segregation algorithms allocate speech regions to participating speakers depending on which speaker dominates in which spectro-temporal region. The proposed method is a different approach to speech segregation, in that it separates the participating speaker streams rather than decide in the favor of the dominating speaker. The algorithm depends on a Lease-Squares fitting approach to model the speech mixture as a sum of complex exponentials. The algorithm gives results that are better than an existent algorithm when tested on the same task. The performance on a different database yielded good segregation results, even for Target-to-Masker ratios as low as -15 dB. The algorithm has immense promise for improvement and practical implementation.

Index Terms **—** speech segregation, speech separation, co-channel speech, monaural speech, auditory scene analysis, Target-to-Masker Ratio

1. INTRODUCTION

Speech signals are seldom available in pure form for speech processing applications, and are often corrupted by acoustic interference like background noise, distortion, simultaneous speech from another speaker etc. In such scenarios, it becomes necessary to first separate the speech from the background. In particular, the task of separating overlapping speech from multiple speakers, called Speech Segregation, is especially challenging since it involves separating signals having very similar statistic and acoustic characteristics. The segregation problem has attracted immense research effort in the past two decades [1], more so for the case when the mixture is available only from a single channel and multi-channel approaches cannot be used. This single-channel situation is called the co-channel or monaural speech segregation problem. In this paper, we focus on and discuss the two-speaker co-channel segregation problem.

The ideal approach towards segregation should identify the perceptually salient features of the participating streams, and preserve all those features during segregation. Recent methods have achieved this goal to an extent [1]. However, they do not completely reconstruct all portions of the participating streams. The long-term goal of recovering speech streams as close as possible to the original signals is yet to be achieved, and remains a bottleneck in speech processing applications. In this paper, we promote a new philosophy towards segregation, and propose an algorithm that performs better than a state-of-the-art algorithm [2].

2. ANALYSIS OF EXISTENT ALGORITHMS

In both model-based (c.f. [3]) and feature-based (c.f. [2]) approaches towards segregation, the speech signal is first decomposed into a number of channels using a filter-bank over all time frames, giving a collection of Time-Frequency Units (TFUs). Features are extracted from each TFU for analysis. In model-based approaches, each TFU is assigned to the speaker whose model has the maximum likelihood of generating its feature. In feature-based approaches, the features of the TFU are analyzed to identify which of the sources the features match better with, and the TFUs are accordingly assigned to that source. In both cases, each TFU is assigned to one of the two speakers using some decision criteria. The signals within the TFUs corresponding to each source are then used as-is to reconstruct that source.

These approaches assume that TFUs contain energy from one of the two speakers. This assumption causes "leakage errors" and "missed speech" during reconstruction, since most TFUs typically carry speech from both speakers. This is explained with the aid of Fig. 1, which shows a speech utterance from the Cooke database [4] containing speech from a male (A) and female (B) speaker. Panel 1 shows the spectrogram of the mixture. Panel 2 shows the spectrogram of speech from A, with the speech-present regions highlighted in blue. Panel 3 shows the same information for B in red. As can be seen, there exist some TFUs where the mixture contains energy from *both* speakers – a violation of the assumption. Panel 4 shows TFUs where speaker A has greater energy than (dominates) B, and Panel 5 shows TFUs where B dominates A. Current algorithms aim at recovering these latter two profiles accurately – the so-called Ideal T-F (ITF_{DOM}) maps. Following this, TFUs dominated by speaker A are used to reconstruct the speech of A, and those dominated by B are used to reconstruct the speech of B. The reconstructed speech of A will consequently contain leaked speech from B

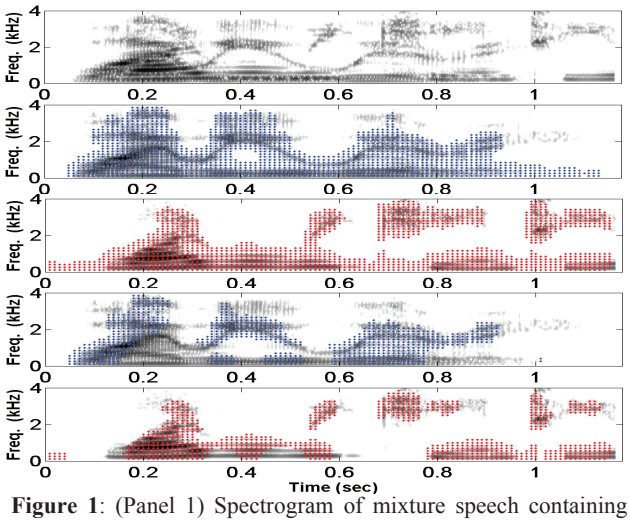

Figure 1: (Panel 1) Spectrogram of mixture speech containing speakers A and B (Panel 2) Spectrogram of A, with non-silent TFUs in blue (Panel 3) Spectrogram of B, with non-silent TFUs in red (Panel 4) Spectrogram of A, with the dominant TFUs in blue (Panel 5) Spectrogram of B, with the dominant TFUs in red.

(Leakage Error) since some TFUs actually contain energy from both speakers, and the reconstructed speech of B would lack some speech (Missed Speech). This error is not due to the inability to recover the ITF_{DOM} masks accurately. Rather, it is due to the inability of the ITF_{DOM} masks to describe the original signals completely. An implication of this fact is that in cases where the Target-to-Masker Ratio is less than 0 dB, a significant portion of the target would be dominated by the masker and will be irrecoverable. For good segregation, it is necessary to distribute the signal content among both speakers. We thus aim to estimate all non-silent regions of both utterances – the Complete Ideal T-F mask (ITF_{COM}). Indeed, Missing Feature Theory [5] has the aim of bridging the gap between the ITF_{DOM} and ITF_{COM}. We estimate the ITF_{COM} directly using a model that recovers *both* the signals in the mixture.

3. SEGREGATION : A LEAST SQUARES PROBLEM

The algorithm performs segregation by modeling each TFU as a combination of complex sinusoids which are harmonics of the pitch frequencies of the speakers. Thus, it requires a multi-pitch detector [6]. Our algorithm is different from previous models using sinusoids [7] that are spectrum-based and estimate only the amplitudes of the sinusoids. Such algorithms are susceptible to the effects of windowing. Our algorithm, on the other hand, directly models the time series and thus is not affected by window parameters. It can also estimate *both* the amplitudes & phases of the sinusoids.

Consider a Fourier Series representation of a stationary periodic signal $x[n]$,

$$x[n] = \sum_{i=1}^{N} (\alpha_i^+ \exp(j\omega_0 in) + \alpha_i^- \exp(-j\omega_0 in)) \qquad (1)$$

where $i = 1, 2, .., N$ are the N harmonics of the fundamental frequency ω_0, α_i is the amplitude of the i^{th} harmonic and can be complex, and the + and – represent the α_i for positive and negative frequencies. Then, for a sequence x[n], if we want to estimate the unknown amplitudes α_i^+ & α_i^-, it can be done by using $M > 2N$ different values of $x[n]$. Substituting n = 1, 2, …, M in equation (1) and obtaining M equations in the N unknown coefficients, we have in vector form

$$\underline{x} = [V^+ \; V^-]\,\underline{\alpha} = A\,\underline{\alpha} \qquad (2)$$

where

$\underline{x} = [x[1]\; x[2]\; \ldots\; x[M]]^T$, $\underline{\alpha} = [\alpha_1^+ \; \alpha_2^+ \ldots \alpha_N^+ \; \alpha_1^- \; \alpha_2^- \ldots \alpha_N]^T$,

$V^+ = [\,\underline{v}[1]\; \underline{v}[2]\; \ldots\; \underline{v}[N]\,]$, $V^- = (V^+)^*$ and

$\underline{v}[k] = [\exp(j\omega_0 1k)\;\; \exp(j\omega_0 2k)\;\; \ldots \exp(j\omega_0 Mk)]^T$

where the superscript * represents conjugation and ^{T} represents transpose. If $M > 2N$, this gives an over-determined system of equations. The least square error solution of (2) is $\underline{\alpha} = A^P\underline{x}$, where $A^P = (A^H A)^{-1}A^H$ is the pseudo-inverse of A. Observe that the matrix V^+ (and V^-) is composed of columns which form a set of basis functions or signals. The coefficients can therefore also be found by the Gram-Schmidt procedure. Since $\underline{\alpha}$ is complex, *both* the amplitudes *and* phases of the complex exponentials are estimated. We build on this principle by dividing the input into TFUs and assuming that the signal in each TFU is stationary (i.e. the coefficients $\underline{\alpha}$ are the same within a TFU). The input mixture signal is passed through an analysis filter-bank that decomposes the input into a number of channels. Analysis is done on a frame-wise basis with overlapping frames, yielding a number of TFUs. For each TFU, if the energy is below a threshold, the TFU is labeled silent and not analyzed further. For all non-silent TFUs, the pitch is used to estimate the two streams as described below.

3.1. Segregation of Voiced-Voiced Speech

For a given TFU, if the pitch of both speakers is non-zero, the mixture signal being analyzed is the sum of two (quasi) periodic signals, $s_A[n]$ and $s_B[n]$. Let the mixture signal be $x_{TF}[n]$ and the pitch values be ω_A and ω_B. Then:

$$x_{TF}[n] = s_{A,TF}[n] + s_{B,TF}[n] = \sum_{i=1}^{N_A} (\alpha_i^+ \exp(j\omega_A in) + \alpha_i^- \exp(-j\omega_A in)) + \sum_{k=1}^{N_B} (\beta_k^+ \exp(j\omega_B kn) + \beta_k^- \exp(-j\omega_B kn))$$

giving, in vector form: $\underline{x}_{TF}[n] = \underline{v}_A{}^T[n]\,\underline{\alpha} + \underline{v}_B{}^T[n]\,\underline{\beta}$

where the set of parameters $\{\underline{\alpha}\} = [\,\alpha_1^+ \; \alpha_2^+ \ldots \alpha_{N1}^+ \; \alpha_1^- \;\; \alpha_2^- \ldots \alpha_{N1}^-\,]^T$ corresponds to the voiced component of speaker A and the set of parameters $\{\underline{\beta}\} = [\,\beta_1^+ \; \beta_2^+ \ldots \beta_{N2}^+ \; \beta_1^- \;\; \beta_2^- \ldots \beta_{N2}^-\,]^T$ corresponds to the voiced component of B. Here, N_A & N_B are the number of harmonics existing between 0 and $F_S/2$, where F_S is the sampling frequency. By choosing the length of a T-F unit as $M > 2(N_A + N_B)$, $\{\underline{\alpha}\}$ & $\{\underline{\beta}\}$ are obtained as the least squares solution to the set of equations:

$$\underline{x} = [V_A^+ \; V_A^- \; V_B^+ \; V_B^-][\,\underline{\alpha}^T \;\; \underline{\beta}^T\,]^T = V\underline{\gamma}$$

Having obtained the estimate $\{\underline{\alpha}'\}$, which defines $s_A[n]$, and $\{\underline{\beta}'\}$, which defines $s_B[n]$, we can reconstruct both the signals $s_{A,TF}[n]$ & $s_{B,TF}[n]$ that composed the mixture as:

$$s_{A,TF}[n]' = v_A{}^T[n]\alpha' = \sum_{i=1}^{N_A}(\alpha'^{+}_i \exp(j\omega_A in) + \alpha'^{-}_i \exp(-j\omega_A in))$$

$$s_{B,TF}[n]' = v_B{}^T[n]\beta' = \sum_{k=1}^{N_B}(\beta'^{+}_k \exp(j\omega_B kn) + \beta'^{-}_k \exp(-j\omega_B kn))$$

3.2. Presence of Unvoiced Speech

For a given TFU, if one of the speakers (say B) is unvoiced ($\omega_B = 0$) then the observed mixture signal can be modeled as

$$x_{TF}[n] = s_{A,TF}[n] + s_{B,TF}[n] =$$
$$\sum_{i=1}^{N_A}(\alpha_i^{+}\exp(j\omega_A in) + \alpha_i^{-}\exp(-j\omega_A in)) + w[n] = \underline{v}_A{}^T[n]\underline{\alpha} + w[n]$$

where $w[n]$ represents a noise source. This can expressed in vector form for a window length $M > 2N_A$ as

$$\underline{x} = [\mathrm{V_A}^{+}\ \mathrm{V_A}^{-}]\ \underline{\alpha} + \underline{w}$$

where $\underline{w}$ is a noise vector. Since it is hard to find a set of basis functions for the noise component $\underline{w}$, we pursue an alternate solution for $\underline{\alpha}$. Let E be the energy of the signal $x_{TF}[n]$. The energy of $s_{A,TF}[n]$ is equal to $E_A = ||\underline{\alpha}||^2$. We thus have a constraint that $||\underline{\alpha}||^2 \le E$. The LS fitting problem now becomes a constrained minimization problem:

$$\textit{minimize over } \underline{\alpha} : ||\ [\mathrm{V_A}^{+}\ \mathrm{V_A}^{-}]\ \underline{\alpha} - \underline{x}||^2 \textit{ subject to } ||\underline{\alpha}||^2 \le E$$

This is a convex-cost, convex-constraint minimization problem and Second Order Cone Programming (SOCP) yields a solution to this problem [8]. Once the estimate of the coefficients $\{\underline{\alpha}\}$ is obtained, the source signal of A is reconstructed exactly as above. Since the energy of the unvoiced source E_B is the difference between E and E_A, one way of obtaining the unvoiced signal B is to generate a White Gaussian Noise of variance E_B. This procedure serves to generate fricative-like sounds in all regions of unvoiced speech, even for frames where the true signal was a stop. Reconstructing the actual obstruent in the TFU (i.e., fricative, stop etc.) is a subject of future research. Since the proposed algorithm depends on a pitch-based model, currently the algorithm does not have the provision to deal with TFUs where both speakers are unvoiced. Segregation of such TFUs is another future goal.

Since we use overlapping frames in our analysis, the reconstructed signals from successive TFUs cannot be appended to each other. The estimated signals from TFUs are added across channels to reconstruct the estimate for each frame first. The stream of each speaker is then reconstructed across frames using the overlap-add method.

4. EVALUATION

Fig. 2 shows the performance of the algorithm in separating speakers from the same mixture as in Fig. 1. Performance in recovering both the ITF_{COM} and ITF_{DOM} masks are shown. The Estimated Complete TF mask (ETF_{COM}) of each speaker is obtained by finding all non-silent regions of the *reconstructed* speech. It is seen that most of the speech-present region has been well captured for *both* speakers. Similarly, for *both* speakers, the Estimated Dominated TF mask (ETF_{DOM}), obtained in a similar way as the ITF_{DOM} but from the reconstructed streams, matches the ITF_{DOM} well.

The algorithm was quantitatively evaluated using two different databases. The pitch tracks needed to perform the estimation were extracted from the original signals using ESPS Wavesurfer. We compare the performance of the proposed algorithm with the Hu-Wang (HW) algorithm in [2] on the same task. The Cooke database [4] is used, and the task is to recover the voiced utterance in the presence of 3 distinct masker speech signals, *n7*, *n8* and *n9*. The metric of comparison is the Percentage of Energy Loss ($\mathrm{P_{ELD}}$) and the Percentage of Noise Residue ($\mathrm{P_{NRD}}$), defined as follows:

$$P_{ELD} = \frac{\text{Total energy of TFUs with } \{\mathrm{ITF_{DOM}} = 1, \mathrm{ETF_{DOM}} = 0\}}{\text{Total energy of TFUs with } \mathrm{ITF_{DOM}} = 1}$$

$$P_{NRD} = \frac{\text{Total energy of TFUs with } \{\mathrm{ITF_{DOM}} = 0, \mathrm{ETF_{DOM}} = 1\}}{\text{Total energy of TFUs with } \mathrm{ITF_{DOM}} = 1}$$

$\mathrm{P_{ELD}}$ is the relative energy present in the ITF_{DOM} but missing from the ETF_{DOM} and $\mathrm{P_{NRD}}$ is the relative energy absent in ITF_{DOM} but detected as present in ETF_{DOM}. Ideally, both these figures must be low. We also present the SNRs of the reconstructed target signals, defined as the ratio between the energy of the target signal to that of the reconstruction error:

$$SNR = \frac{\sum_n s^2[n]}{\sum_n (s[n] - \hat{s}[n])^2}$$

Here the SNR is calculated using the entire original signal. Finally, as proposed by us, the ITF_{DOM} does not provide complete information about the goodness of reconstruction and thus we also present the values $\mathrm{P_{ELC}}$ and $\mathrm{P_{NRC}}$ which are obtained by substituting the ITF_{COM} mask for ITF_{DOM} in the

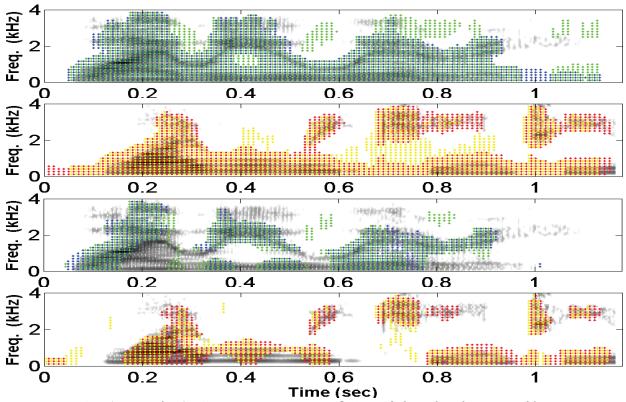

Figure 2: (Panel 1) Spectrogram of A with Ideal non-silent TFUs in blue and Estimated non-silent TFUs in green (Panel 2) Spectrogram of B with Ideal non-silent TFUs in red and Estimated non-silent TFUs in yellow (Panel 3) Spectrogram of A with Ideal dominant TFUs in blue and Estimated dominant TFUs in green (Panel 4) Spectrogram of B with Ideal dominant TFUs marked in red and Estimated dominant TFUs in yellow.

equations above. Table 1 gives our results on this database:

	Hu-Wang [6]		Proposed Algorithm				
Masker	P_{ELD}	P_{NRD}	P_{ELD}	P_{NRD}	P_{ELC}	P_{NRC}	SNR
n7	2.01	2.25	2.01	1.56	0.06	0.003	10.29
n8	1.16	0.65	6.44	0.18	0.15	0.001	11.68
n9	17.8	14.22	3.77	12.83	0.15	0.002	7.00

Table 1: Results on the Cooke database [4]. All values are expressed in percentages, except for the SNR which is expressed in dB. Compare : columns 1 & 3; columns 2 & 4.

The proposed system performs segregation better than the Hu-Wang algorithm for almost all masker signals with lower values of P_{ELD} & P_{NRD}. Also, the P_{ELC} & P_{NRC} values are very low indicating that most of the speech-present regions are well preserved in the reconstruction, and the missed regions or falsely identified regions are low-energy TFUs. The SNR of the reconstructed signals also confirm that the quality of reconstruction by our algorithm is good.

The algorithm is also evaluated on a larger database consisting of synthesized mixtures from the TIMIT database. The task is to recover one of the signals in the mixture (the target). Three different test sets were synthesized – {Male, Male} (MM), {Female, Male} (FM) and {Female, Female} (FF). In the FM set, half the mixtures had a male target and the other half a female target. 200 utterances were generated for each set at seven different Target-to-Masker ratios (TMRs): -15 dB, -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, 15dB, giving a total of 21 sets. The performance was tested across a wide range of TMRs in order to examine how well the theory of shared TFUs is supported by our algorithm. If the energy is accurately being divided between the two sources, our algorithm should perform well even at low TMRs.

The results are shown in Fig. 4. The Error Loss and Noise Residue trends are consistent across gender sets. On average, performance is best for the FF population and worst for the MM population. This might be attributed to the fact that the pitch values of males are closer to each other than those of females, making estimation harder. The values of P_{ELC} & P_{NRC} are significantly lower than that of P_{ELD} & P_{ELD} (compare the scales of the two plots). This is due to the "shared" TFU concept, which results in fewer units with significant energy being missed or falsely detected. The quality of reconstruction can be seen from the SNR plot of the recovered signals. Even at low TMRs the SNR of the recovered signal is above 0 dB. For all TMRs upto 10 dB, the algorithm gives an SNR greater than the TMR and yields speech that is more usable than the original mixture. This result demonstrates that even at very low TMRs, the streams can be pulled apart and good quality segregation is indeed possible. Audio samples of the mixtures and reconstructed signals can be found at [9].

5. CONCLUSIONS

We have presented an approach to the speech segregation problem which emphasizes analyzing the speech regions in a way as to pull apart the participating speakers rather than decide in favor of the dominating one. The algorithm gives separation results comparable to or better than an existent algorithm and shows good performance at very low TMRs. Some aspects of our future research will be (1) segregation when both speakers are unvoiced, (2) segregation when the pitch frequencies of both speakers are very close, (3) improving segregation in very low TMR and in the presence of noise, (4) extension to more than two speakers.

11. REFERENCES

[1] D. L. Wang & G. Brown, eds. (2006), *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley/IEEE Press, Hoboken, NJ, 2006.

[2] G. Hu & D. L. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control*, edited by E. Hansler & G. Schmidt, Springer, Heidelberg, Germany, pp. 485–515., 2006.

[3] G. J. Brown & M. P. Cooke, "Computational auditory scene analysis", Comput. Speech and Language, 8, pp. 297-336, 1994.

[4] M. P. Cooke, *Modeling Auditory Processing and Organisation*, Cambridge University Press, Cambridge, UK, 1993.

[5] M. Cooke & G. Brown, "Separating simultaneous sound sources: issues, challenges and models," in *Fundamentals of speech synthesios and speech recognition* Edited by E. Keller, John Wiley & Sons, pp. 295-312, 1994.

[6] S. Vishnubhotla & C. Espy-Wilson, "An Algorithm for Multi-Pitch Tracking in Co-Channel Speech", Proceedings of Interspeech 2008, Melbourne, Australia, 2008.

[7] T.F Quatieri & R.G. Danisewicz,, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," IEEE Transactions on Acoustics, Speech and Signal Processing, , vol.38, no.1, pp.56-69, Jan 1990.

[8] S. Boyd & L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[9] http://www.glue.umd.edu/~srikanth/SS_samples/

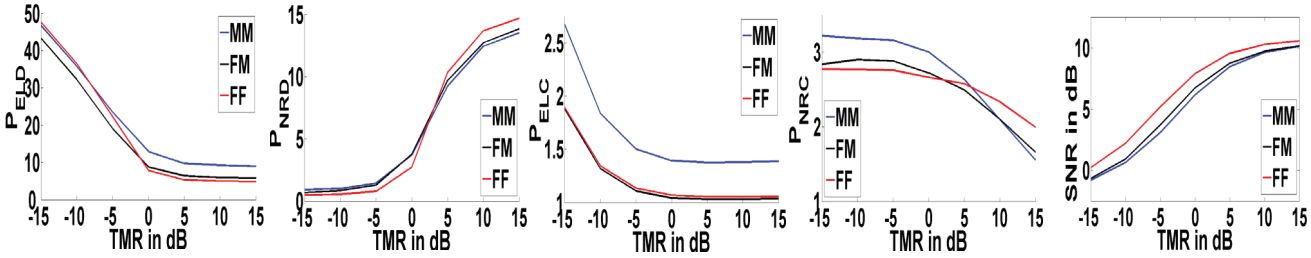

Figure 4: Performance of the algorithm on the TIMIT database. The percentage of energy loss, P_{EL}, and percentage of noise residue, P_{NR}, are shown for both the ITF_{DOM} and ITF_{COM} at different TMRs. The SNR of the reconstructed signal is also shown.