Speech Segregation from Co-channel Mixtures

Srikanth Vishnubhotla and Carol Espy-Wilson
Speech Communication Laboratory, Institute for Systems Research & ECE Department, University of Maryland, College Park, USA
Email : { srikanth ; espy } @ umd . edu

THE PROBLEM

We ***have*** a single channel (microphone/telephone etc.) recording containing overlapping speech from multiple speakers, potentially with noise in the background

We ***want*** to automatically extract the clean speech from noise, and separate the different speaker "streams" from the mixture (recording) for further speech applications

THE MOTIVATION

Technology : Most speech processing applications (ASR, SID) rely on the assumption of "clean speech" and their performance degrades rapidly with increasing interference and noise

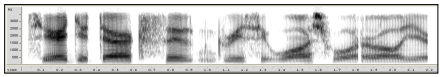
Health Services : Users of Hearing Aids and Cochlear Implants have trouble when faced with "cocktail-party" situations, and the signal processing inside these devices would benefit from speech enhancement and segregation algorithms

Single Sensor : It may be impractical to use multiple sensors

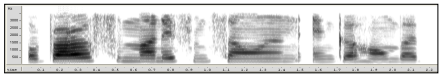
THE DIFFICULTY

Overlapping speech sounds have similar acoustic & statistical characteristics. Therefore, it is difficult to come up with an algorithm that can separate out overlapping speech-like sounds and assign the separated streams to the appropriate speaker

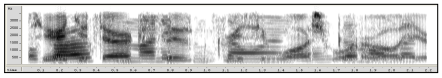
AN EXAMPLE



Spectrogram of Speech from Speaker A



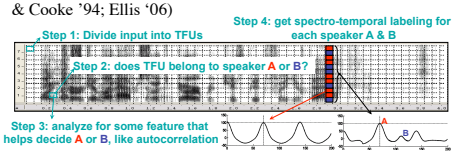
Spectrogram of Speech from Speaker B



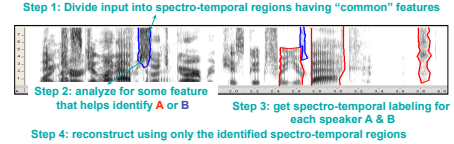
Spectrogram of Mixture Speech containing Speakers A & B

ANALYSIS OF CURRENT APPROACHES

Time-Frequency Unit (TFU) level (Wang & Brown '06; Brown & Cooke '94; Ellis '06)



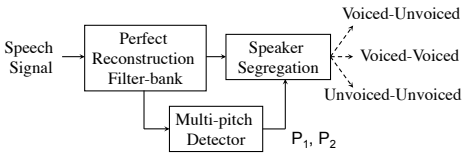
Spectrogram level (Barker, Cooke & Ellis '05; Hu & Wang 07)



Issue : these algorithms generate "hard" masks, with each spectro-temporal region being assigned to only one speaker

Proposed Approach : share the energy in each TFU between ***both*** the speakers. We use harmonicity as a cue to tease apart the contributions of the two speakers in each TFU.

THE PROPOSED SPEECH SEGREGATION SYSTEM

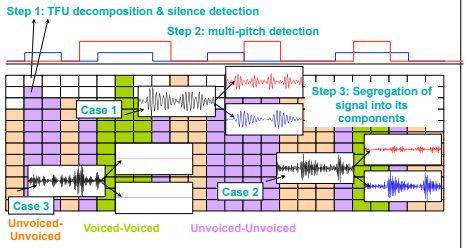


The following are the main stages:

Step 1 : obtain a TFU representation using a PR filter-bank, and find the silent and non-silent (speech-present) regions

Step 2: obtain the pitch of both speakers using a multi-pitch detector (Vishnubhotla & Espy-Wilson, '08)

Step 3 : perform segregation of the signal within each TFU using the pitch values for that time instant, obtained from Step 2. Depending on the two pitch values, this action falls into one of three possible cases to be described next.



Let $x_{TF}[n]$ represent the speech signal within a particular TFU.

Case 1 : Voiced-Voiced (Pitch Frequencies ω_1 & ω_2)

Since $x_{TF}[n]$ is (quasi) periodic, it can be written as

$$x_{TF}[n] = \sum_{i=1}^{N} \alpha_i^{+} \exp(j\omega_i i n) + \sum_{i=1}^{N} \alpha_i^{-} \exp(-j\omega_i i n) + \sum_{k=1}^{P} \beta_k^{+} \exp(j\omega_2 k n) + \sum_{k=1}^{P} \beta_k^{-} \exp(-j\omega_2 k n)$$

and for $n = 1, 2, \ldots, W$ together, can be written in vector form as

$$\underline{x}_{TF} = \begin{bmatrix} V_{CA} & V_{SA} \end{bmatrix} \begin{bmatrix} \underline{\alpha}^{+} \\ \underline{\alpha}^{-} \end{bmatrix} + \begin{bmatrix} V_{CB} & V_{SB} \end{bmatrix} \begin{bmatrix} \underline{\beta}^{+} \\ \underline{\beta}^{-} \end{bmatrix} = [V][\underline{\gamma}]$$

where V is known by construction and $\underline{x}_{TF}$ is observed. The unknown coefficient $\underline{\gamma}$ and thus the individual contributions of the speakers are found using Least Squares Estimation (LSE) :

$\underline{\tilde{\gamma}} = V^{+} \underline{x}_{TF}$ where $\underline{\gamma}^{H} = \begin{bmatrix} \underline{\alpha}^{+} & \underline{\alpha}^{-} & \underline{\beta}^{+} & \underline{\beta}^{-} \end{bmatrix}^{H}$

From this, we get the signals of the two speakers in that TFU:

$$\underline{\tilde{s}^{(1)}}_{TF} = \begin{bmatrix} V_{CA} & V_{SA} \end{bmatrix} \begin{bmatrix} \underline{\tilde{\alpha}}^{+} \\ \underline{\tilde{\alpha}}^{-} \end{bmatrix} \qquad \underline{\tilde{s}^{(2)}}_{TF} = \begin{bmatrix} V_{CB} & V_{SB} \end{bmatrix} \begin{bmatrix} \underline{\tilde{\beta}}^{+} \\ \underline{\tilde{\beta}}^{-} \end{bmatrix}$$

Case 2 : Voiced-Unvoiced (Pitch Frequency ω_1)

In this case, $x_{TF}[n]$ can be written as

$$x_{TF}[n] = \sum_{i=1}^{N} \alpha_i^{+} \exp(j\omega_i i n) + \sum_{i=1}^{N} \alpha_i^{-} \exp(-j\omega_i i n) + (w_R[n] + j w_C[n])$$

and for n = 1, 2, …, W together, can be written in vector form as

$$\underline{x}_{TF} = \begin{bmatrix} V_{CA} & V_{SA} \end{bmatrix} \begin{bmatrix} \underline{\alpha}^{+} \\ \underline{\alpha}^{-} \end{bmatrix} + \underline{w} = V\underline{\gamma} + \underline{w}$$

where V and $\underline{x}_{TF}$ are known. Since LSE is sensitive to the strength of the noise $\underline{w}$, we estimate $\underline{\gamma}$ differently. We assume that most of the energy in the TFU comes from the voiced speaker and the energy of the unvoiced speaker is the minimum possible. We then have a constrained minimization problem:

$$\underline{\tilde{\gamma}} = \min_{\underline{\gamma}} \| \underline{x}_{TF} - V\underline{\gamma} \|^2 \text{ subject to } E(\underline{s^{(1)}}_{TF}) \le E(\underline{x}_{TF})$$

where $E(\underline{x}_{TF}) = \| \underline{x}_{TF} \|^2$ and $E(\underline{s^{(1)}}_{TF}) = \| \underline{s^{(1)}}_{TF} \|^2 = \| \underline{\gamma} \|^2$

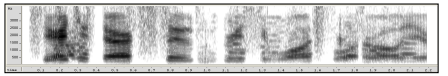
The individual contributions are then obtained by generating the voiced signal as in the previous case, and the unvoiced signal using white noise of variance equal to its power:

$$\underline{\tilde{s}^{(1)}}_{TF} = \begin{bmatrix} V_{CA} & V_{SA} \end{bmatrix} \begin{bmatrix} \underline{\tilde{\alpha}}^{+} \\ \underline{\tilde{\alpha}}^{-} \end{bmatrix} \qquad \underline{\tilde{s}^{(2)}}_{TF} = wgn(0, E(\underline{x}_{TF}) - E(\underline{s^{(1)}}_{TF}))$$

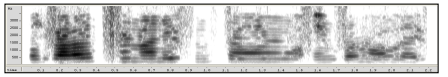
Case 3 : Unvoiced-Unvoiced (Both Pitch Frequencies = 0)

The model currently does not account for unvoiced-unvoiced regions. This will also be a goal of future research. For now, we replace such regions by silence, in the reconstructed streams.

PERFORMANCE OF THE SEGREGATION SYSTEM

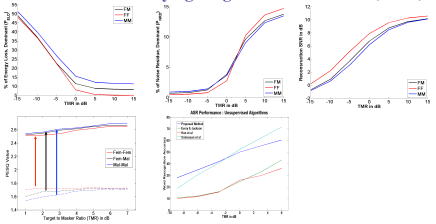


Spectrogram of Reconstructed Speech from Speaker A



Spectrogram of Reconstructed Speech from Speaker B

Performance with varying Target-to-Masker Ratio (TMR):



Even at low TMRs, the perceptual quality of reconstruction is good. On the task of automatic speech recognition, the proposed algorithm ranks highest for the low TMR cases.

REFERENCES

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

Brown, G & Cooke, M P (1994), "Computational auditory scene analysis", Comput. Speech and Language, 8, pp. 297-336.

Ellis, D (2006) "Model-Based Scene Analysis" in "Computational Auditory Scene Analysis: Principles, Algorithms, and Applications", Wang, D L & Brown, G, eds., Wiley/IEEE Press, pp. 115-146.

Barker, J, Cooke, M P & Ellis, D (2005), "Decoding speech in the presence of other sources", Speech Communication, vol. 45, no. 4, pp. 5 – 25.

Hu, G & Wang, D L (2007), "Auditory segmentation based on onset and offset analysis", IEEE Trans. on Audio, Speech, and Language Processing, vol. 15, pp. 396-405.

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

ACKNOWLEDGEMENTS

This research was supported by NSF Grant # BCS-0519256 and ONR grant # N00014-07-M-0349