S^2LAP: Source Separation and Localization with Array Processing

Mark Bryk
The Cooper Union
New York, NY

Christopher Curro
The Cooper Union
New York, NY

David Katz
The Cooper Union
New York, NY

Abstract

A system is proposed for separation and localization of speech sources for which searchable transcriptions can be made. This system is to be implemented with a 32 element microphone array, a data acquisition device, and a digital computer.

I. Introduction

In a crowded and noisy room, humans are remarkably able to focus on the voice of a single speaker. This ability is named the cocktail party effect and much research has been done to reproduce a similar effect in electronic systems using microphones and signal processing to mirror the ears and the brain [1]. For example, researchers have used multiple microphones, since binaural hearing improves one's ability to display the effect; humans with only one ear have much more difficulty separating sources [2]. It has been proven that in a system with M microphones and N sources, if $M \geq N$ then the sources can be completely separated [3]. However, in this case, the microphones need to be strategically placed throughout the room.

We propose a device which can produce labeled transcriptions of any conversation occurring in a room. It will stand beside the wall or hang from the ceiling, skirting the issue of inconveniently located microphones. The device will create audio channels for each individual speaker whether they are seated next to each other or apart. These noiseless and separated streams will then be ready for speech to text conversion, and the transcriptions stored in an easily navigable database.

Such a device has many useful applications. Transcribed meetings - whether for business offices, journalist's interviews, courthouses, or government agencies - save time and enforce transparency. Real-time transcriptions on a large display can also service the hearing-impaired and the deaf.

II. System Description

The system consists of several components. The physical device is a microphone array connected to a computer via a data acquisition device. The data acquisition software will output the individual audio channels from each microphone. The signal processing begins with an angle of arrival estimation to determine the amount and location of the sources, relative to the array. Once the sources have been identified, beamforming filters will be constructed to make separate channels for each estimated angle of arrival. If it is determined that two sources are approximately collocated, then source separation techniques, such as independent component analysis, will be used. After source separation, the independent speech channels will be piped to Google for speech to text conversion. Speakers will be classified by their location in space, the spectral characteristics of their speech, and the word content of their speech using natural language processing. The transcription will be stored in a database and can be easily searched.

A. Microphone Arrays

A 32 element microphone array is the first element in the transcription system. The array can be subdivided to maximally support the source separation algorithms, post beamforming. Each microphone will be approximately isotropic, i.e., have an approximately uniform spherical gain. Furthermore, the microphones will have maximally flat frequency responses across the audio band. Each microphone will need to be calibrated to minimize the differences in output characteristics between them.

B. Data Acquisition

A data acquisition system will be used to capture input from the microphones. The data acquisition software will output 32 independent channels from each microphone synchronized sample-wise.

C. Angle of Arrival Estimation

The angular locations of the speech sources will be estimated by locating the maxima in the multiple signal classification (MUSIC) spectrum. The MUSIC spectrum is defined as:

$$S_{\text{MUSIC}}(\theta) = \frac{1}{\mathbf{s}^{\text{H}}(\theta)\mathbf{P}_v\mathbf{s}(\theta)} \tag{1}$$

where θ is the electrical angle, $\mathbf{s}(\theta)$ is a steering vector, and $\mathbf{P}_v$ is the projection matrix onto the noise subspace. The peaks in the MUSIC spectrum correspond to unique physical angles of arrival, as long as the array element spacing is less than half a wavelength. The steering vectors used for the beamforming process are calculated by this angle of arrival estimation process.

D. Beamforming

Beamforming is a technique which allows an array of isotropic antennas to function as a single highly directionalized antenna [4]. The output from an antenna array is of the form:

$$A = \begin{bmatrix}\mathbf{u}(1) & \cdots & \mathbf{u}(L)\end{bmatrix} \tag{2}$$

where:

$$\mathbf{u}(t) = \begin{bmatrix} u(\mathbf{r}_1, t) \\ \vdots \\ u(\mathbf{r}_M, t) \end{bmatrix} \tag{3}$$

In a spatial sampling system, such as an antenna array, the columns of the A matrix, the $\mathbf{u}(t)$ vectors, are discrete-time complex exponentials, or in the case of a multiple source system, sums of discrete-time complex exponentials. Because of this fact, finite impulse response (FIR) filters can be constructed to only accept signals impinging from a particular direction. The number of taps in the vector $\mathbf{w}$ for the FIR filter is equal to the number of antenna elements. For the case of multiple sources, generalized sidelobe cancelers can be constructed to remove interference from a specified direction [5]. This can be solved as a problem of linear constraints:

$$\mathbf{C}^{\mathrm{H}}\mathbf{w} = \mathbf{g} \tag{4}$$

where $\mathbf{C}$ is the constraint matrix, and $\mathbf{g}$ is the gain vector. In the case of two sources - one target and one interferer - the equation becomes:

$$\left[\mathbf{s}(\theta_0)\ \mathbf{s}(\theta_1)\right]^{\mathrm{H}}\mathbf{w} = \begin{bmatrix}1\\0\end{bmatrix} \tag{5}$$

The optimal solution to the problem can be found with an adaptive algorithm. By filtering with the $\mathbf{w}$ vector, individual channels for speakers based on angle of arrival can be constructed.

The above discussion operates under the assumption that the signals incident on the microphone array are narrowband. In order to accommodate realistic wideband signals, orthogonal filter banks will be used to separate the wideband signals into narrow subbands. That is each microphone channel will be decomposed into several subbands. The subbands from across the array will be processed together. The beamformed responses for each subband will then be combined together again to reconstruct the wideband signal.

E. Source Separation

Common algorithms for blind source separation, such as independent component analysis, work under the assumption that the number of channels is greater than or equal to the number of speakers. To ensure this relationship, the microphone array will be subdivided into two smaller arrays if two speakers are determined to be within the main beam width of the array. Each sub-array will compute beamforming coefficients and spatially filter their signals independently. This process sacrifices main beam width but allows the separation of speakers by other algorithms.

F. Speech to Text

A variety of different open source tools are available for speech to text conversion; the performance of these platforms is dependent on the clarity of the speaker's words and the amount of noise in the surroundings. Thus, successful separation of voices will result in more accurate text, allowing the final transcript to serve as an evaluation metric for the system.

We plan on using Google's API, which also provides a confidence metric with the predicted text. It determines the likelihood that a specific sequence of words or phrases make sense together. By combining these two metrics - accuracy and confidence - of text predictions over a large set of spoken data, we can comprehensively evaluate our signal processing success.

G. Speaker Classification

Speaker identification can be performed via a Gaussian Mixture Model (GMM) classifier [6]. Mel-frequency cepstral coefficients (MFCCs) can be computed for each frame. These coefficients represent the spectral shape of a frame of audio. Frames that overlap by 50% can be used to capture subtle changes in voice. Over a large window these coefficients can be used to generate a GMM model for each speaker. To identify a speaker, the likelihood of a sequence of frames must be computed using the models of each stored speaker. The speaker will be identified as

the one with the maximum likelihood. However, if all likelihoods fall below a certain threshold, then it will be assumed that the current speaker is a new speaker, and a new model will be generated for him or her.

Location will also play a large role in successful speaker classification. Our system will operate under the assumption that a speaker will not be moving faster than a walking pace. Thus, as a speaker traverses the room, their position will be tracked via angle of arrival. When classifying speakers, the location information of each utterance can be used as well, allowing us to create a model both of the speaker and their position over time.

Further speaker classification can be done manually, via a web interface. There, users can both correct classification errors and also attribute a speaker from a conversation to an existing speaker ID. It is possible that the computer will have separated one speaker into two, or combined two speakers into one. The user can manually combine or separate these speakers, respectively. Also, the user can give a name to any speaker, thus associating a speaker with a pre-existing speaker ID and making the data more navigable. We also want to provide the option to train the system for specific speakers. This will allow the most frequent users of the system to have a higher chance of being properly detected.

H. Database, Search and Navigation

The transcriptions will be made available through a database which can be easily searched via a web interface. The data will be shown as a list of lines, sorted by timestamp, grouped by conversation, with the speaker's name listed before each line. A user will be able to nagivate the transcriptions through a variety of filters. A user will be able sort transcriptions by speaker ID, conversation ID, or any other relevant ID. A user will also be able to search through the entire database for specific keywords.

REFERENCES

[1] A. Bronkhorst, "The cocktail party phenomenon: A review on speech intelligibilty in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, pp. 117–128, 2000.

[2] M. Hawley, R. Litovsky, and J. Culling, "The benefit of binaural hearing in a cocktail party: effect of location and type of interferer," *Journal of the American Accoustics Society*, vol. 115, no. 2, pp. 833–843, 2004.

[3] S. of Sparse and N.-S. M. in Source Separation, *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, July 2005.

[4] M. Richards, *Fundamentals of Radar Signal Processing*. McGraw Hill, 2005.

[5] S. Haykin, *Adaptive Filter Theory*, 5th ed. Pearson, 2014.

[6] H. Lu, A. J. B. Brush, B. Priyantha, A. K. Karlson, and J. Liu, "Speakersense: Energy efficient unobtrusive speaker identification on mobile phones," *Pervasive Computing*, vol. 6696, pp. 188–205, 2011.