THE RELATION BETWEEN SPEECH SEGMENT SELECTIVITY AND SOURCE LOCALIZATION ACCURACY

Parham Aarabi, Alborz Mahdavi

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto
10 Kings College Road
Ontario, Canada, M5S 3G4

ABSTRACT

An experimental analysis of the relation between speech signal segment power and the source direction-of-arrival estimation accuracy is conducted. A total of 10 different speakers, including both male and female speakers, totaling to approximately 2 hours of speech are used to analyze the performance of the Phase Transform, the Maximum Likelihood, and the Unfiltered Cross Correlation time-delay estimation techniques. For female speakers, it is determined that the Phase Transform technique has a lower percentage of anomalies and a lower direction-of-arrival root mean-square error (DOA RMSE). Conversely, for male speakers, it is determined that the Unfiltered Cross Correlation has a lower percentage of anomalies although the Phase Transform has a lower DOA RMSE. The spatial distribution of the errors as well as the speech segment power relation to the errors are also presented.

1. INTRODUCTION

The localization of sound sources has been arises in a variety of different applications, including speech recognition [13, 3], intelligent rooms [4, 1], and various other applications [2]. Many different techniques for sound localization have been proposed in the past. These include, among others, Multiple Signal Classification [11], Maximum Likelihood [12], and Time Delay of Arrival (TDOA) based techniques [8, 5, 9]. Each technique has its advantages and disadvantages, but TDOA-based techniques have recently gained popularity due to their simplicity and efficiency.

The various sound localization techniques utilize a short sound segment (about 10-20 ms long) obtained by an array of two or more microphones. In practice, the suitability of a sound segment must always be questioned since some sound segments may contain the speech of the to-be-localized source while other segments may just contain noise. This paper experimentally analyzes a simple variance threshold method for selecting suitable sound segments and compares the performance of different variance thresholds as well as different TDOA estimation techniques using 10 different speakers including 7 male and 3 female speakers each talking from a total of 10 spatial positions. TDOA-based sound localization involves computing the time-delays of the sound signal between a pair of microphones. This estimated time-delay constrains the sound source location to a hyperbola in 2-D. Multiple microphone pairs result in multiple hyperbolae whose intersection ideally corresponds to the sound source location.

2. TIME DIFFERENCE OF ARRIVAL BASED SOUND LOCALIZATION

Using Time Difference of Arrivals (TDOAs) between pairs of microphones is a prominent technique for sound source localization [8, 5, 1, 10, 6, 2]. Since each TDOA constrains the sound source location to a hyperbola in 3-D, the presence of 3 microphone pairs is theoretically enough to pinpoint the sound source location. This location estimate, however, is very much dependent on the accuracy of the TDOA estimates, which itself depends on the acoustics of the environment, the signal-to-noise ratio (SNR), the type of speaker, the speech segment size, and the TDOA estimation algorithm.

TDOA estimation algorithms can be described as follows:

Assuming we have a sound signal of interest $s(t)$ that is arriving at the first microphone, we can describe the signals obtained by a pair of microphones as:

$$x_1(t) = h_1(t) * s(t) + n_1(t) \quad (1)$$

$$x_2(t) = h_2(t) * s(t - \tau) + n_2(t) \quad (2)$$

where $x_1(t)$ and $x_2(t)$ are the signals received by the first and second microphones, respectively, $h_1(t)$ and $h_2(t)$ are the room impulse responses, $n_1(t)$ and $n_2(t)$ are the noise signals of the respective microphones, and τ is the TDOA

between the two microphones corresponding to the main sound signal.

Our goal is to estimate the TDOA by observing a small (about 10 to 20 ms) segment of the signals received by the microphones. One method of accomplishing this is by using the general frequency-domain cross correlation function [8]:

$$\tilde{\tau} = \arg\max_{\beta} \int_{-\infty}^{\infty} X_1(\omega) e^{-j\omega\beta} \overline{X_2(\omega)} d\omega \qquad (3)$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of $x_1(t)$ and $x_2(t)$ respectively, and $\overline{X_2(\omega)}$ is the complex conjugate of $X_2(\omega)$.

Recent work on TDOA estimation has shown that in practical situations, a frequency weighing function can improve estimation accuracy [8, 6, 2, 10]. In low signal-to-noise environments, assuming that the additive noise signals $n_1(t)$ and $n_2(t)$ are stationary and that their spectra are known a-priori, the Maximum Likelihood weighing function [7] can be applied resulting in the following Maximum Likelihood TDOA estimate:

$$\tilde{\tau}_{ML} = \arg\max_{\beta} \int_{-\infty}^{\infty} \frac{\frac{X_1(\omega) e^{-j\omega\beta} \overline{X_2(\omega)}}{|X_1(\omega)||X_2(\omega)|}}{\frac{|N_1(\omega)|^2}{|X_1(\omega)|^2} + \frac{|N_2(\omega)|^2}{|X_2(\omega)|^2}} d\omega \qquad (4)$$

In the case of reverberant rooms, the Maximum Likelihood TDOA weighing function usually does not perform very well [3, 8]. In such situations, the Phase Transform weighing function results in a more accurate TDOA estimate [8]. This estimate can be written as:

$$\tilde{\tau}_{PHAT} = \arg\max_{\beta} \int_{-\infty}^{\infty} \frac{X_1(\omega) e^{-j\omega\beta} \overline{X_2(\omega)}}{|X_1(\omega)|\,|X_2(\omega)|} d\omega \qquad (5)$$

It should be noted that Direction of Arrival (DOA) estimates can be directly obtained from the TDOA estimates. If a pair of microphones are d apart, the DOA θ which measures the estimated angle of the source from the perpendicular bisector of the line between the microphones can be defined as:

$$\theta = \arcsin(\frac{\tilde{\tau} v}{d}) \qquad (6)$$

where v is the velocity of sound in air and $\tilde{\tau}$ is the TDOA estimate.

3. SPEECH SEGMENT VALIDITY

In the TDOA estimation process, the validity of a certain segment must be analyzed. For example, a segment may not contain any signal of interest (very low SNR) thereby resulting in an inaccurate TDOA estimate. In such a situation, it is better to specify that this segment is invalid and to not associate equal validity with its result and the result of a segment with a much higher SNR. However, this requires the prior estimation of the SNR followed by an SNR dependent prefiltering of the segments. In this paper, the prefiltering involves either concluding that a segment is valid (has SNR above a threshold) or invalid (has an unacceptable SNR). The estimation of the SNR can be outlines as follows:

Similar to the Maximum Likelihood situation, if we assume that the additive noise sources are stationary so that their power spectra can be estimated from prior segments where the main signal was known to be silent, then we can define the Normalized Variance as:

$$R = \frac{\int_{-\infty}^{\infty} |X_1(\omega)|^2 d\omega}{\int_{-\infty}^{\infty} |N_1(\omega)|^2 d\omega} + \frac{\int_{-\infty}^{\infty} |X_2(\omega)|^2 d\omega}{\int_{-\infty}^{\infty} |N_2(\omega)|^2 d\omega} \qquad (7)$$

The Normalized Variance R can be used as an indication of the validity of an individual segment. In this paper, the effects of choosing different variance thresholds on the TDOA accuracy are experimentally analyzed.

4. EXPERIMENTAL RESULTS

In order to analyze the TDOA estimation accuracy of the different techniques and its relation to the type or location of the speaker, an experiment was conducted in a 7m by 5m room with the microphones and speakers being placed as shown in Figure 1. The total of 10 speakers spoke at each of the ten spatial locations, totaling to approximately 2 hours of speech. The 2 hours of speech were then decomposed into 10ms segments and the Maximum Likelihood (ML), Phase Transform (PT), and the Unfiltered Cross Correlation (UCC) techniques were used to generate TDOA estimates for each segment. The variance of the segments were also recorded and used as a metric for the validity of that segment. If the variance was above a Normalized Variance Threshold (NVT), then the segment's TDOA estimates were taken into account. If the variance did not surpass the NVT, then its TDOA estimates were not considered as if the segment never existed.

In reporting the localization errors, two different parameters, the percent anomalies of the DOA and the root mean-square error of the DOAs that are not anomalies are reported. This is typical of reporting TDOA estimation accuracy [8]. Percent anomalies in this paper are defined as the percentage of direction of arrival estimates that are within $\pm 5^o$ of the true direction of the sound source.

4.1. TDOA Estimation Accuracy of Male Speakers

Figure 2 illustrates the percent anomalies in the source localizations for the 7 male speakers in relation to the NVT. As the NVT increases, the percentage of anomalies decreases. This comes at the cost of a smaller number of segments that are available for processing. In Figure 2, it is clear that the

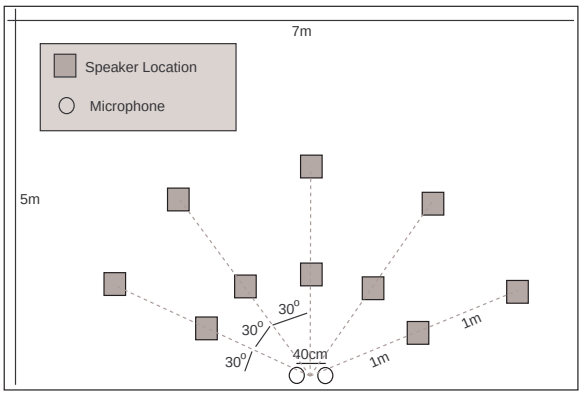

Fig. 1. A 5m by 7m environment used for the sound localization experiments. A total of 10 different spatial locations were used for each speaker.

UCC technique surpasses the other two at almost all different values of the NVT.

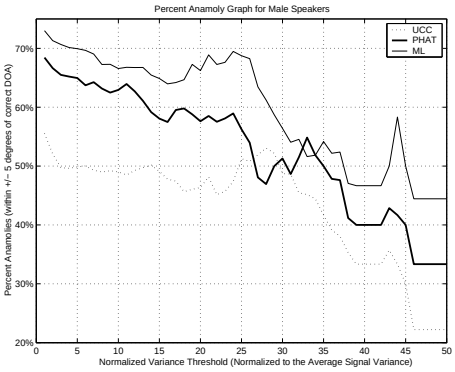

Fig. 2. Relation between the direction-of-arrival of sound anomalies and the normalized variance threshold for male speakers.

Figure 3 shows the Root Mean-Square Error (RMSE) of the DOAs that are not anomalies. Here, the PHAT estimates have a slightly lower RMSE which becomes substantially lower than the other two techniques at high NVTs.

4.2. TDOA Estimation Accuracy of Female Speakers

A similar set of results are computed for the 3 female speakers. The relation between percent anomalies and the NVT are shown in Figure 4.

In Figure 4, the Phase Transform performs better than the other two techniques. The Phase Transform also results in a lower RMSE for the female speakers as shown in Figure 5.

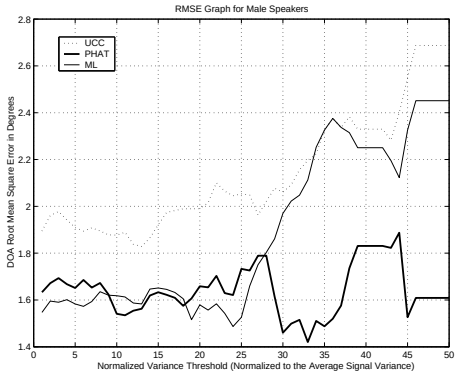

Fig. 3. Relation between the root mean-square error of the direction-of-arrival of sound and the normalized variance threshold for male speakers.

Fig. 4. Relation between the direction-of-arrival of sound anomalies and the normalized variance threshold for female speakers.

4.3. Spatial Distribution of DOA Errors

The results shown previously for male and female speakers are combined and shown for the 10 different spatial positions in Figure 6. Figure 6 illustrates the relation between the spatial position of the speaker and the DOA RMSE. Interestingly, the PHAT technique achieves a lower error for the speaker locations further away from the microphone pair than the locations closer to it. This is in contrast with the other techniques, for the locations further away from the microphone pair have a higher error than the closer locations.

5. CONCLUSIONS

The results in this paper illustrate the importance of NVTs for successful sound source localizations. It is intuitively clear that higher NVTs would yield a lower localization error due to their selectivity. This is experimentally shown with the exception of some of the DOA RMSE errors, whose

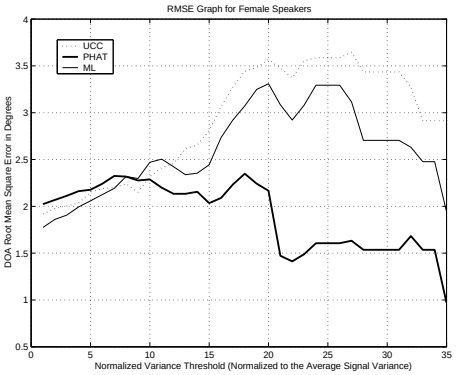

Fig. 5. Relation between the root mean-square error of the direction-of-arrival of sound and the normalized variance threshold for male speakers.

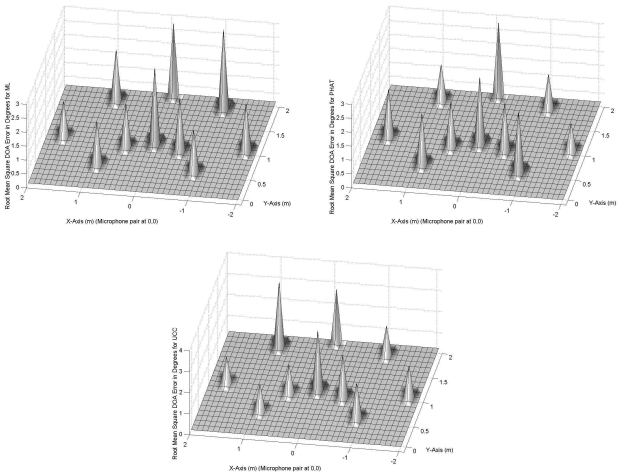

Fig. 6. Spatial distribution of the root mean-square DOA errors for the 10 different spatial positions. This is obtained from all available data from both male and female speakers using an NVT of 10.

increase with increased NVTs can be attributed to a shrinking number of available segments and therefore increased fluctuations in the mean of the results.

Overall, the PHAT technique outperformed the other techniques in TDOA estimation accuracy. This can be attributed to the nature of the environment, where a relatively high signal-to-noise ratio (approximately 15dB) and an environment with strong reverberations (a domininant reverberation time of about 0.03s) made it suitable for the PHAT technique, and less suitable for the UCC or the ML techniques.

6. REFERENCES

[1] P. Aarabi. Multi-sense artificial awareness, June 1998. M.A.Sc. Thesis, Department of Electrical and Computer Engineering, University of Toronto.

[2] P. Aarabi. *The Integration and Localization of Distributed Sensor Arrays*. PhD thesis, Stanford University, May 2001.

[3] P. Aarabi. Robust sound localization using the formant phase transform. In *Proceedings of the 5th IEEE Workshop on Nonlinear Signals and Information Processing*, June 2001.

[4] P. Aarabi and S. Zaky. Integrated vision and sound localization. In *Proceedings of the 3rd International Conference on Information Fusion*, July 2000.

[5] P. Aarabi and S. Zaky. Iterative spatial probability based sound localization. In *Proceedings of the 4th World Multiconference on Circuits, Systems, Computers, and Communications*, July 2000.

[6] M.S. Brandstein. *A Framework for Speech Source Localization Using Sensor Arrays*. PhD thesis, Brown University, May 1995.

[7] M.S. Brandstein, J. Adcock, and H. Silverman. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer, Speech, and Language*, 9:153–169, 1995.

[8] M.S. Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, May 1996.

[9] C. H. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24(4):320–327, August 1976.

[10] D. Rabinkin. A DSP implementation of source location using microphone arrays. In *131st meeting of the Acoustical Society of America*, May 1996.

[11] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, AP-34:276–280, March 1986.

[12] H. Watanabe, M. Suzuki, N. Nagai, and N. Miki. A method for maximum likelihood bearing estimation without nonlinear maximization. *Transactions of the Institute of Electronics, Information, and Communication Engineers*, J72A, 8:303–308, August 1989.

[13] B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice-Hall, New Jersey, 1985.