

CO-CHANNEL SPEAKER SEGMENT SEPARATION

Brett Y Smolenski and Robert E. Yantorno
Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, Pa 19122-6077, USA
bsmolens@astro.temple.edu, robert.yantorno@temple.edu
http://www.temple.edu/speech_lab

Daniel S. Benincasa and Stanley J. Wenndt
Air Force Research Laboratory/IFEC, 32 Brooks Rd. Rome NY 13441-4514, USA
danb@rl.af.mil, wenndts@rl.af.mil

ABSTRACT

A novel approach to co-channel speaker separation is presented here. The technique uses the statistical properties of combinations of high Target-to-Interferer Ratio (TIR) speech segments, which were extracted from a 0 dB overall TIR co-channel utterance. The problem is broken down into making three simpler decisions. First, closed-set speaker identification technology is used on combinations of high TIR speech segments to determine which speakers are generating the co-channel speech. Next, the proportion of segments belonging to each speaker is estimated using a bimodal model. Lastly, a maximum likelihood decision is made as to which two combinations of segments best represent the two speakers. Using this approach at least one of the speakers could readily be identified when the speaker contributed a segment that was 160 ms or more in length. Once the speakers were determined, greater than 90% reliable speaker separation was obtained.

1. INTRODUCTION

The traditional approach to co-channel speech processing has been to enhance the target speech, to attenuate the interfering speech, or to both enhance the target speech while attenuating the interfering speech. Thirty years of these approaches has produced limited results. However, recently a novel approach to co-channel speech processing has been proposed using the concept of usable speech.

Within a co-channel utterance, where both speakers are contributing the same overall energy, there exists several segments of speech where one of the speakers is 20 dB or more above the other speaker. It has been shown that when the target speaker is at least 20 dB greater than the interfering speaker, 80% reliable identification of the target speaker can be obtained [1]. Hence, these segments with a high Target-to-Interferer Ratio (TIR) may be considered usable with respect to speaker identification.

Recent advances in co-channel speech processing have produced several usable speech measures [2] [3] [4]. These measures have high correlation with TIR of co-channel speech. Such measures are necessary to determine usability in an operational environment, since a priori knowledge of the TIR probably will not be available.

Since there would exist usable speech segments of both speakers, there needs to be a way of separating what segments belonged to which speaker. A reliable speaker identification system would be one method that could accomplish this task.

However, a significant challenge with using these “usable” speech segments is that the speaker identification system generally does not perform well with segments of speech shorter than 350 ms [5]. Unfortunately most segments of usable speech are less than 350 ms.

To increase the reliability of the speaker ID system, with respect to segment length, combinations of the usable segments can be used. Theoretically one could test all of the possible combinations of usable speech segments, then the two best scores of combinations should represent the two speakers. However, this approach does not yield good results due to confounding from the other speakers in the training set.

2. METHOD

To reduce confounding, the decision process is broken down into three simpler decisions. First, the speaker ID system is used on combinations of usable speech segments to determine which speakers are generating the co-channel speech. The idea behind this technique is to take advantage of the fact that, given a large sample of usable segments, the two highest frequencies of a histogram of the top scores produced by the speaker ID system represent the two speakers. Generalizations can readily be made when more than two people are talking.

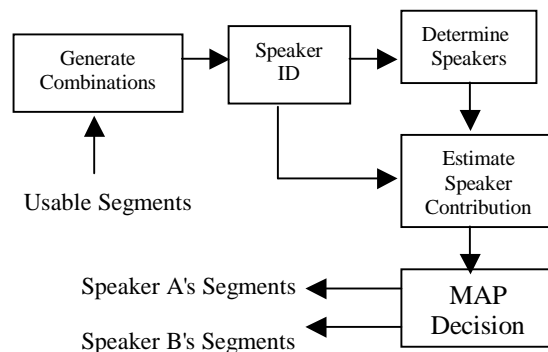


Figure 1. System Block Diagram

The success of this technique relies on being able to obtain enough reliable scores from the speaker identification system. This amounts to being able to obtain enough long segments. Fortunately segments greater than the necessary three 40 ms frames in length are common in co-channel speech [6].

The second phase of this process is to determine the likely proportion of segments contributed by each speaker. This involves generating a histogram of all 38 speakers in the speaker identification system’s training set and fitting it to a theoretical

bimodal distribution. The individual means of each term of this distribution are then computed. Since we know the total number of segments k , the mean values of each component in the distribution can be used as an estimate of the proportion of frames contributed by each speaker [7].

Since two speakers produce the usable segments, one would expect to have a bimodal distribution of the scores. The two modes would represent the centers of mass of the two speakers.

Each of the combinations in these pairs would have two scores, produced by the speaker identification system, associated with them. These scores would be with respect to the two speakers determined in the first part of the process discussed above. These scores are then adjusted by multiplying them by the arithmetic mean of the spectral flatness measure of the frames contained in the combination.

Next, a weighted average is taken of the scores of each segment contained in the combinations of disjoint pairs. This average is computed by first multiplying the score of each segment by the number of frames contained in it, adding the segments together, and then dividing by the total number of frames contained in the two combinations.

Finally, the disjoint pair that had the smallest weighted score was selected as the pair that represented speaker A and speaker B. This decision in effect separates the segments.

To obtain the usable speech segments, several pairs of spoken sentences were taken from the TIMIT database. The longer file in each pair was truncated to make both files the same length and the files were then combined at 0 dB overall TIR to form the co-channel utterance.

It should be noted that in an operational environment it is highly unlikely that two speakers would be talking over each other during every utterance. In addition, each utterance would not be exactly the same length or TIR. The reason for this approach was to capture the worst possible scenario, with respect to both speakers, that one could reasonably expect in a co-channel environment.

Once the co-channel utterance was formed it was broken down into 40 ms frames with no overlap. The usable segments, contiguous frames that had greater than 20 dB TIR, were then extracted. These parameters were later used to obtain a better threshold in the final decision process [8].

Next, the usable segments, the individual frames that comprised the segments, and the individual utterances that formed the co-channel data were passed through the speaker identification system. Thirty-eight speakers were in the systems training set, which included the two talkers forming the co-channel utterance. The training set consisted of five sentences for each speaker from the TIMIT database. Lastly, the scores obtained from the speaker identification system were then processed with the separation system discussed above.

To examine any variability that may exist between male-over-male and female-over-female co-channel speech, the above procedure was performed with 12 pairs of female-over-male utterances, as well as, 11 pairs of both male-over-male and female-over-female speech. Hence, each female that was paired with a female was also paired with a male and vice versa.

To control the variability between the dialect regions, four speakers, two male and two female, were randomly selected from each of the six major dialect regions. This approach follows the Latin Square method of experiment design for controlling two possible sources of error [9]. Research has shown that, provided the utterances are the same length, the characteristics of usable speech depend more on the speaker than on what is being said [10].

3. RESULTS

Figure 1 below shows two histograms with each of the 38 speakers in the training set being the bins. The figure on the left corresponds to results of the top, lowest, scores for all the pairs of combinations. The figure on the right corresponds to results of the top two scores.

As one can see from the histograms shown below, speaker 13 has almost twice as many low scores than any of the other speakers in the training set. In addition, taking the top two scores doesn't seem to produce better results, but the difference between the two can be useful.

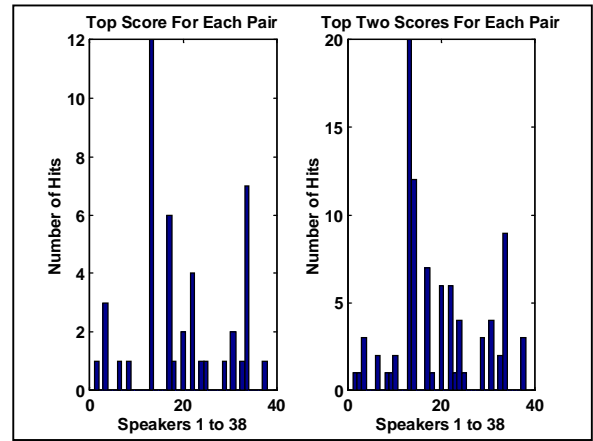


Figure 2. Histograms of top scores for all possible pairs of usable segments

This information can be used to derive a level of significance for the decision. The situation is equivalent to flipping an unfair coin 45 times and considering the probability of a specified number of hits/heads i.e. the binomial law. Hence, from the above example, we can reject the null hypothesis and accept the hypothesis that speaker 13 is one of the speakers at the 99% confidence level.

This research has found that in order for this decision to be valid, the speaker has to contribute at least one segment that is 4 or more frames in length. This is why only speaker 13 is prominent in the histogram. Fortunately, current research

indicates that there is roughly a 50% chance that an utterance will contain such a frame. Thus, after a few utterances, one can reliably decide which two speakers are creating the co-channel speech.

By considering histograms for the top scores of combinations of three or more one can obtain a high level of significance. If pairs didn't show a high enough level of significance one would try combinations of three, since the only way to reduce the probability of making a type-I error is to increase the sample size.

The next step in this process was to estimate the proportion of segments contributed by each speaker. Figure 2 below illustrates how this was accomplished.

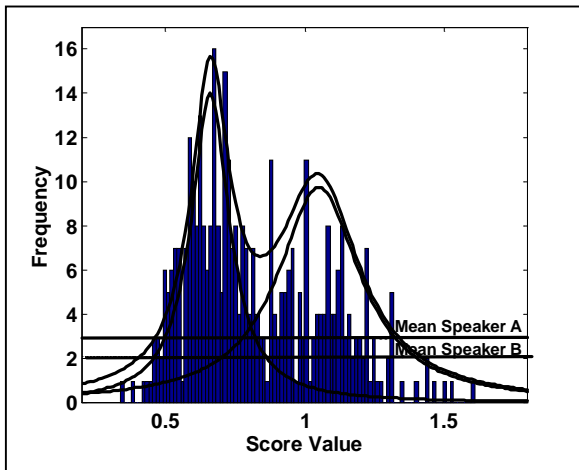


Figure 3. Histogram of top scores for all combinations of three usable segments

The authors believe there are two likely reasons for the success of this modal. First, since the scores are basically Euclidean distances, it would seem reasonable that the concentration of their values would fall by an inverse square law. Second, this technique is, in principal, very similar to using the Gaussian Mixture Model in that samples are being taken from multiple distribution with different means [11]. However, since there are only 38 samples/scores a more appropriate mixture would make use of Student's *t* distribution, which is an inverse square law [12]. This distribution is commonly used for small sample sizes and asymptotically approaches the normal distribution as the sample size increases.

For each of the 34 experiments performed this technique correctly identified the proportions of segments from each speaker. The 1 out of 10 segments that were incorrectly identified were in part due to the use of the speaker ID system in deciding which of the disjoint pairs of combinations best represented each speaker.

4. CONCLUSIONS

The above results demonstrate that, under the appropriate conditions, a speaker identification system can be used to identify as well as separate usable speech segments. Speaker identification does, however, necessitate having adequate training data, which in an operational environment may not be available.

The necessary conditions for using this technique include; at least one of the segments should have a high reliability, which depends on its normalized score, length, and energy; and the segments should be contiguous; the number of frames separating each segment should be as small as possible.

The entire system could be used for situations where there are more than two people talking. Probably the most significant downsides to using this approach are its computational complexity and the fact that at this time it cannot run in real time. It is always necessary to first collect and process at least 10 usable segments first in order to make reliable estimates.

5. FURTHER RESEARCH

An improved understanding of how the characteristics of usable segments and their combinations effect the scoring of the identification system will lead to choosing a better decision threshold.

The speaker identifications system could possibly be further optimized for speech segments. Further, a specific type of closed-set speaker identification system was used to separate the segments. It is conceivably possible that other classifiers could be used and could even be more effective.

Perhaps the most promising approach that should be explored would be to include in the weighting scheme the value of the TIR or, equivalently, the value of a usable speech measure. It seems intuitive that there would exist different degrees of usability as opposed to the simple dichotomy of good and bad. For example, frames slightly below the threshold, which were adjacent to a segment meeting the threshold, would most likely be useful.

Further, the various artifices that comprise the process can have useful applications by themselves. The system that estimates the proportion of segments contributed by each speaker could conceivably serve as a speaker counter that would count the number of speakers in multi co-channel speech, such as, counting number of people talking in a room.

ACKNOWLEDGEMENT

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number F30602-00-1-0517. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

DISCLAIMER

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory, or the U.S. Government.

6. REFERENCES

- [1] Yantorno, R. E., "Co-Channel speech and speaker identification study", Final report for Summer Research Faculty Program, Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1998.
- [2] Krishnamachari, K. R., Yantorno, R. E., Benincasa D. S., and Wenndt, S. J., "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions." IEEE International Symposium. Intelligent. Sig. Process. and Comm. Sys., Nov. 2000.
- [3] Krishnamachari, K. R., Yantorno, R. E., Lovekin J. M., Benincasa, D. S., and Wenndt, S. J., "Use of Local Kurtosis Measure for Spotting Usable Speech Segments in Co-channel Speech." ICASSP 2001, pp: 649-652, May 2001.
- [4] Lovekin, J., Krishnamachari, K. R., Yantorno, R. E., Benincasa, D. S., and Wenndt, S. J., "Adjacent Pitch Period Comparison (APPC) as a Usability Measure of Speech Segments Under Co-channel Conditions". IEEE International Symposium on Intelligent Signal Processing and Communication Systems, November, 2001.
- [5] Lovekin, J., Yantorno, R. E., Benincasa, S., Wenndt, S., and Huggins, M., "Developing Usable Speech Criteria for Speaker Identification", ICASSP 2001, pp: 421-424, May 2001.
- [6] Yantorno, R.E., "Fusion – The Next Step In Usable Speech Detection", Final report for Summer Research Faculty Program, Research Laboratory AFRL/IF, Speech Processing Lab, Rome Labs, New York, 2001.
- [7] Jackson, B. W., and Thoro, D., "Applied Combinatorics with Problem Solving". Addison-Wesley: Reading, MA, 1990.
- [8] Barkat, M., "Signal Detection and Estimation". Artech House: Boston, MA, 1991
- [9] Sheskin, D., "Statistical Tests and Experimental Design: A Guidebook". Gardner Press: New York, NY, 1984.
- [10] Ricart, R., "Speaker Identification Technology". RL-TR-95-275: Final Technical Report, Sponsored by AFRL/IF, Rome, NY. 1996.
- [11] Hamming, R. W., "The Art of Probability: For Scientists and Engineers". Addison-Wesley: Reading, MA, 1991.
- [12] Petrucci, J. D., Nandram, B., and Chen, M., "Applied Statistics for Engineers and Scientists". Prentice Hall: Upper Saddle River, NJ. 1999.