

# Effects of co-channel speech on speaker identification

Robert E. Yantorno<sup>a</sup>, Daniel S. Benincasa<sup>b</sup> and Stanley J. Wenndt<sup>b</sup>

<sup>a</sup>Temple University/ECE Dept., 12<sup>th</sup> & Norris Streets, Philadelphia, Pa 19122-6077, USA

<sup>b</sup>Air Force Research Laboratory/IFEC, 32 Brooks Rd., Rome NY 13441-4514, USA

## ABSTRACT

Past studies have shown that speaker identification (SID) algorithms that utilize LPC cepstral feature and a vector quantization classifier can be sensitive to changes in environmental conditions. Many experiments have examined the effects of noise on the LPC cepstral feature. This work studies the effects of co-channel speech on a speaker identification (SID) system. It has been found that co-channel interference will degrade the performance of a speaker identification system, but not significantly when compared to the effects of wideband noise on an SID system. Our results show that when the interfering speaker is modeled as one of the speakers within the training set, it has less of an effect on the performance of an SID system than when the interfering speaker is outside the set of modeled speakers.

**Keywords:** Co-channel Speech, Speaker Identification, Speech Processing

## 1. INTRODUCTION

Speaker identification (SID) techniques have applications in security access, telephone transmissions, forensic science, and situational awareness. In a closed-set SID system, the goal is to identify an unknown speaker from a set of  $M$  possible speakers. Along with the spoken message, good quality speech contains information about the identity of the speaker. The SID task is to measure the information contained in the speech, which is unique to each speaker. This task begins by collecting speech training data to generate speaker models for each speaker. When test data is collected under similar conditions as the training data, and is of good quality, excellent identification performance (over 90% correct identification) can be achieved. However, this is not always the case. In many situations, such as law enforcement and military applications, co-channel speech can adversely corrupt a suspects speech, thereby resulting in poor performance.

Co-channel speech is defined as a speech signal that is a combination of speech from two or more talkers recorded over a single communication channel. Historically, the goal of co-channel research has been to be able to extract the speech of one of the talkers from the co-channel speech. This can be achieved by either enhancing the target speech or suppressing the interfering speech. This co-channel situation has presented a challenge to speech researchers for the past 30 years [1,2,3,5]. When the extracted speech is to be used in an SID system, then one must determine how much and what type of target speech is needed to perform "good" speaker identification, i.e., how much interference is acceptable. Therefore, determining the effect of speaker interference on speaker identification would be of considerable interest. The goal of this research is to better understand the co-channel problem and how it impacts the performance of an SID algorithm.

## 2. CO-CHANNEL INTERFERENCE ON SPEAKER IDENTIFICATION

The goal of co-channel research has been to extract the speech of one of the talkers from the co-channel speech. This can be achieved by either enhancing the target speech or suppressing the interfering speech. The question researchers faced in the past was how to extract the speech of one of the speakers. However, if the final goal, with respect to co-channel speech, is to use it for such things as speaker identification, then there are two approaches that might be used. The first is to identify if co-channel speech is present or absent. If it is absent then one can safely process the speech without worrying about the quality or usability of that speech. The second would be to determine if the co-channel speech might actually be useable for such things as speaker identification. Therefore, when processing any information it would be essential that one know how "good" that information is. This is especially true when one has only a limited amount of data, which is typically the case with military and law enforcement situations.

When two people are talking face-to-face, the speech signal is only one of many signals used to transfer information. However, when only speech is used for communication, as in the case of the telephone, one is much more limited in the

amount of information that is sent. Therefore, it is very important to ensure that incoming speech, which is going to be used for such things as speaker identification, speaker verification, speech recognition or language identification, is reliable so that the results of any of those speech processing systems are also reliable.

Several experiments were conducted. Speech samples were taken from the TIMIT database. The system was trained using 15 male and 15 female speakers. The number of files for training for each speaker was 5. The files were taken from the dialect region 1 (DR1 subdirectory). Different speech files from the same 30 speakers were used for testing. The speaker identification tests were conducted under closed-set conditions. The files used for training were the “SX” prefix speech files and the files used for testing were the “SA” and “SI” prefix speech files. The SID system used an LPC cepstral feature set and a vector quantization classifier.

Previous work has examined the effects of noise on the LPC cepstral feature [4]. If one assumes that wideband noise represents an extreme condition for corrupting speech for identification purpose, then certain conclusions can be drawn about the level, in dB, and the amount of speech (added to the target speech) that might be tolerable before one would observe a significant drop in performance of an SID system on co-channel speech. For example, it has been observed that our SID system can tolerate up to 40% of the target speech signal corrupted with 0dB SNR (wideband noise) with only a slight decrease (of about 15 percent) in percent correct [7].

Two sets of two experiments were conducted. In the first set, the corrupting speech was drawn from one of the speakers of the training and testing data, but was not the same utterance as used for that speaker’s training or testing. These experiments are identified as “closed set” experiments and the results are shown in Figure 1. In the second experiment, the corrupting speech was drawn from speakers outside the training and testing data. These experiments are identified as “open set”, and the results are shown in Figure 2. For both the closed and open set experiments there were four different types of tests: 1.) male speech corrupted by either male or female speech (results are identified as male) and female speech corrupted by either male or female speech (results are identified as female). One major observation that can be made with respect to Figures 1 and 2 is that even with 100% of the speech signal corrupted with an interfering speech signal equal in energy to that of the target signal, there still exists a significant number of correct identifications, i.e., about 40% accuracy.

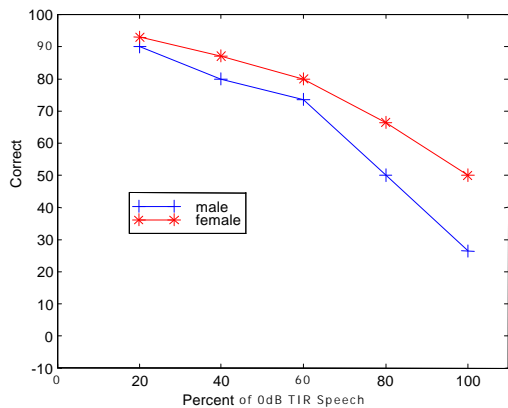


Figure 1(a)

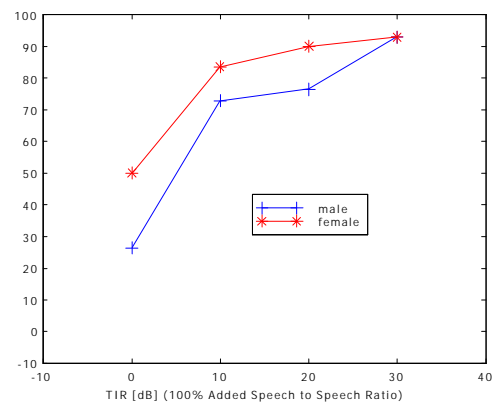


Figure 1(b)

Figure 1. “Closed Set” Speaker Identification Experiments. Figure 1a. Percent Correct versus Percent of 0 dB TIR (Target-to-Interferer Ratio). Figure 1b. – Percent Correct versus TIR in dB of corrupting speech added to speech (100% of speech corrupted by an interfering speech signal).

This result seems reasonable because 0 dB TIR does not spread the energy over the entire utterance, as would be the case of speech corrupted by wideband noise. It is also evident that although there is an almost linear inverse relationship between percent correct and the percent of target speech that has been corrupted (Figure 2a.). The slope is not as steep as one would find with speech corrupted by wideband noise.[11] Also, for the closed set experiments (Figure 1), male

speaker identification appears to be more sensitive to interfering speech than female speaker identification. A smaller effect can be observed with the open set experiments shown in Figure 2. Finally, a comparison is made between the “open” and “closed” set experiments with the results shown in Figure 3. The major observation to be made is that for speaker identification, corrupting speech with speech from a speaker outside the training data tends to have a greater effect on the percent correct than corrupting speech from within the training data. For a speech signal with 100% of the signal corrupted by a speech signal with the same average energy, the percent correct was 40% (for the closed condition) and 35% (for the open condition), or about 5% decrease in percent correct.

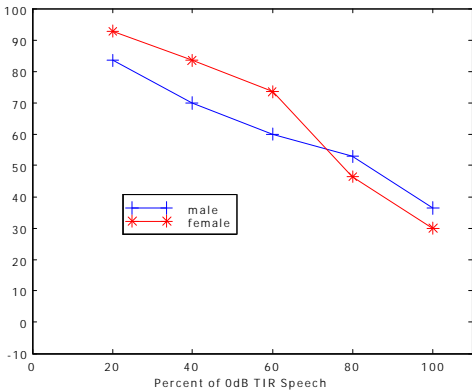


Figure 2 (a)

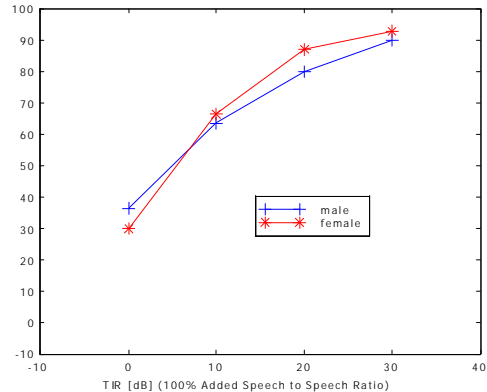


Figure 2 (b)

Figure 2. “Open Set” Speaker Identification Experiments. Figure 2a. Percent Correct versus Percent of 0 dB TIR. Figure 2b. – Percent Correct versus TIR in dB of corrupting speech added to speech (100% of speech corrupted by speech).

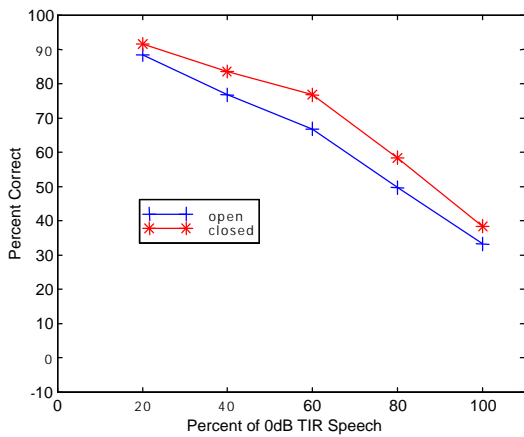


Figure 3 (a)

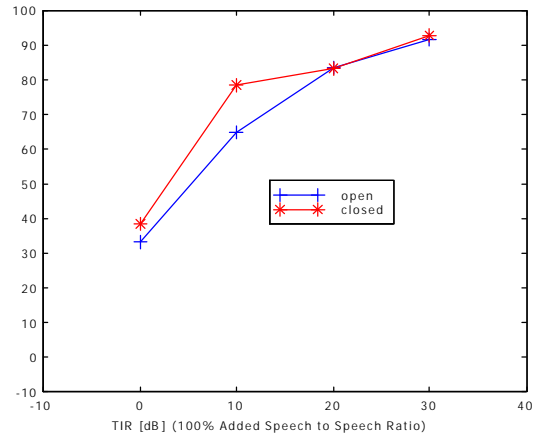


Figure 3 (b)

Figure 3. Comparison of data from “Closed Set” and “Open Set” Speaker Identification Experiments. Figure 3a. Percent Correct versus Percent of 0 dB TIR. Figure 3b. –Percent Correct versus TIR in dB of corrupting speech added to speech (100% of speech corrupted by speech).

It should be noted that Yu & Gish (1993) [8] obtained comparable results for experiments similar to the ones shown in figures 1 and 2. Their goal was to identify either one or both speakers engaged in dialog using speech segments rather than frames and speaker clustering.

### 3. DISCUSSION

It has been shown that co-channel interference can have a significant effect on an SID system. Corrupting speech from a speaker outside the training set tends to have a greater effect on identification when compared to interference from a speaker that has been modeled within the training set. Research needs to be conducted in developing techniques to identify co-channel speech. Since we currently cannot separate speech effectively, we can still extract relevant information from the corrupted speech signal if we know which frames contain just the target speaker and which frames are corrupted. However, to be able to identify co-channel speech on a real-time basis independent of the speakers, may require using unorthodox types of approaches. It seems reasonable that co-channel speech, which is the result of speech corrupted by speech, will not have the same time domain structure as traditional speech, and therefore will not have the same type of phonetic structure as single speaker speech.

### REFERENCES

1. Benincasa, D. S. and Savic, M. I., "Co-channel speaker separation using constrained nonlinear optimization," Proc. IEEE ICASSP, pp:1195-1198, 1997.
2. Meyer, G. F., Plante, F., and Bethommier, "Segregation of concurrent speech with the reassignment spectrum," Proc. IEEE ICASSP, pp:1203-1206, 1997.
3. Morgan, D. P., George, E. B., Lee, L. T., and Kay, S. M., "Co-channel speaker separation," Proc. IEEE ICASSP, pp:828-831, 1995.
4. Openshaw, J. P. and Mason, J. S., "On the limitations of cepstral features in noise," Proc. IEEE ICASSP, pp: II-49-II-52, 1994.
5. Savic, M., Gao, H. and Sorensen, J. S., "Co-channel speaker separation based on maximum-likelihood deconvolution," IEEE ICASSP, pp:I-25-I-28, 1994.
6. Yen, K-C and Zhao, Y., "Co-channel speech separation for robust automatic speech recognition: stability and efficiency," Proc. IEEE ICASSP, pp:859-862, 1997.
7. Yantorno, R. E., "Co-channel Speech and Speaker Identification," Internal AFRL Report, September 1998.
8. Yu, G., and Gish, H., "Identification of speakers engaged in dialog," Proc. IEEE ICASSP, pp:II-383 – II-386, 1993.