

# Developing Usable Speech Criteria for Speaker Identification Technology

*Jereme M. Lovekin, Robert E. Yantorno and Kasturi R. Krishnamachari*  
Temple University/ECE Dept. 1947 N. 12<sup>th</sup> St., Philadelphia, Pa 19122-6077, USA  
jlovekin@temple.edu, ryantorn@nimbus.temple.edu, kkrish01@astro.temple.edu

*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

Recently, a “usable speech” extraction system [1] was proposed to separate co-channel speech into “usable” frames that are minimally corrupted by interfering speech. Studies indicate [2] that a significant amount of co-channel speech can be considered “usable” for speaker identification (SID). Therefore, it is necessary to establish criteria for usable speech frames for SID. Voiced speech, of which usable speech is entirely comprised, is shown to be information rich for SID. In addition, SID accuracy increases as the frame-based Target to Interferer Ratio (TIR) increases when evaluated independently of the amount of available segments. Recent work [3] develops a frame-based Spectral Autocorrelation Ratio (SAR) technique for determining usable frames within co-channel speech. The ability of the SAR method to determine usable frames at various thresholds is examined. This paper investigates the effectiveness of a frame-based usable speech extraction technique for speaker identification.

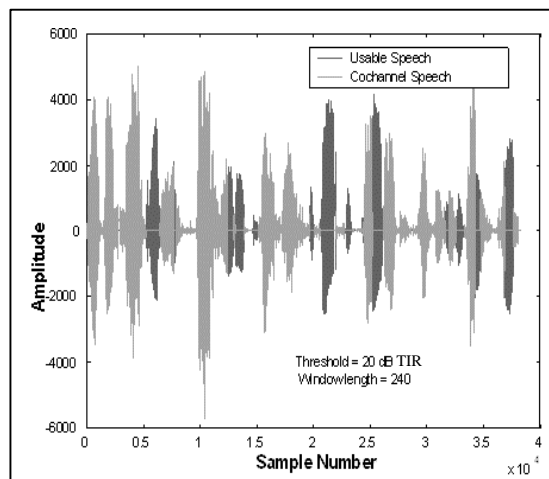
## 1. INTRODUCTION

The usable speech concept is a novel approach to processing co-channel speech, a problem that has challenged the signal processing community for decades. It differs from traditional approaches such as harmonic enhancement and suppression, which attempt to enhance the target speech and attenuate the interfering speech based on harmonic pitch information [4]. Instead, the usable speech approach detects the existence of frames of minimally corrupted speech in a co-channel utterance. Such frames of speech occur when one speaker has much higher energy than another speaker, such as when the target speech is voiced and the interfering speech is unvoiced or silent.

The concept of usability is context dependent. Therefore, it is necessary to define usability based upon its intended application. For example, speech defined as “usable” for speaker identification may not necessarily be usable for speech recognition or voice communication.

For example, with speaker identification, the cepstral features are extracted from the speech signal in order to model a speaker’s voice. Therefore, the speaker’s voice can be modeled based on non-sequential speech information. The speaker identification system can then use the extracted usable frames of speech to determine the identity of a particular speaker, relying on information that is sequence independent. In contrast, for speech recognition, it is necessary to have sequential speech information so that words and sentence structure can be properly identified [5]. Therefore, references to usable speech in this paper will refer to speech usable for speaker identification only.

Usable segments are extracted by detecting the portions of co-channel speech that contain minimal interfering speech from the interfering speaker. Such segments can then be defined as “usable” and sent on for further processing. *Figure 1* shows the original co-channel speech utterance, with approximately 20% usable speech segments (determined by TIR) shown in black, and unusable segments shown in gray.



*Figure 1: Co-channel Speech Utterance. Usable Speech (black) and Unusable Speech (gray)*

## 2. USABLE SPEECH EXTRACTION

Due to the nature of the usable segment extraction system used, only voiced segments will be deemed as

“usable” segments [2]. For this reason, it is expected that the input to the SID system will be a collection of voiced frames from each speaker. Therefore, it is meaningful to extract only voiced frames from the full speaker utterances, and assess the performance of the SID system with these segments to approximate the performance with usable segments. The voiced-only speech is extracted using the Spectral Flatness Method (SFM) [6].

Usable speech frames are also extracted using the frame TIR method, which compares the energy of the target talker and the interfering talker on a frame-by-frame basis. Previous results have shown that the SID can achieve approximately 80% correct identification when the overall TIR is 20 dB [1]. This study expands on that concept by performing a frame based TIR as opposed to an overall TIR. Usable frames of speech are separated and collected into a file for each speaker by calculating the TIR for each frame individually to determine if it exceeds a predetermined threshold.

Finally, usable segments were extracted using the Spectral Autocorrelation Ratio (SAR) method, which takes advantage of the structure of voiced speech in the frequency domain [3]. The SAR method classifies frames of speech as either usable or co-channel. Therefore, the output of the SAR extraction is usable frames. However, the usable frames could correspond to usable speech of either the target speaker or the interfering speaker. In order to determine if the usable frame is from the target speaker or the interfering speaker, the frame TIR at 0 dB is used to assign the frames to either talker. For example, the frames with a positive TIR pertain to one speaker, and those frames with a negative TIR pertain to the other speaker. A comparison can then be made between TIR extracted results and SAR extracted results to determine the ability of the SAR measure to select usable frames.

The original speech data was obtained using 38 speakers from the DARPA TIMIT speech database. The original speech was sampled at 16 kHz, re-sampled to 8 kHz and then low-pass filtered to 3 kHz. Using the three usable speech extraction methods, (SFM, TIR and SAR), usable portions for each speaker were extracted from co-channel utterances. This process resulted in time-compressed segments of data. Because the SID system was intended to process normal speech, a change in performance was expected when using the extracted usable speech.

### 3. SID PERFORMANCE WITH USABLE SPEECH

The SID system involves two steps. First, speakers must be “trained” into the system, so that the system can create a speaker’s model. The system can then be “tested”

with speech from any of the previously trained speakers using different speech samples, which will then compare the given speech to the speaker’s models in an attempt to find a match. In certain situations, full utterances will be available for training, such as when the system is trained or tested in a controlled environment free from co-channel interfering speech. In other situations, such as a field application, only usable segments extracted from co-channel speech may be available. It is also possible that the system could be trained under controlled conditions and tested in the presence of co-channel interference. Therefore, it is necessary to determine performance of SID with usable segments for testing and training both under optimal conditions and in the presence of co-channel corruption. Voiced-only segments extracted at 37 SFM were used for this purpose in place of the actual usable segments. *Table 1* shows the different training and testing situations, with accompanying speaker identification results.

**Table 1: Speaker Identification with Various Voicing States for Training and Testing**

| Training Data | Testing Data | % Correct Speaker ID |
|---------------|--------------|----------------------|
| normal        | normal       | 94.7                 |
| voiced        | voiced       | 77.9                 |
| normal        | voiced       | 75.8                 |
| voiced        | normal       | 35.8                 |
| normal        | unvoiced     | 26.8                 |
| voiced        | unvoiced     | 9.7                  |

The information in *Table 1* is arranged from highest SID accuracy to lowest. The SID accuracy for normal training and testing files is nearly 95%. When voiced only segments were used for training and testing, approximately 80% speaker ID accuracy was achieved. It was realized that less information was available when removing the unvoiced portions of the speech. Correct identification of 75.8% of speakers is attained for the system trained on normal speech and tested on voiced only speech segments, which represents the situation that will most likely be encountered when testing usable segments. It is evident from the above experiment that the extracted voiced segments contain much of the information that is useful for speaker identification.

As previously stated, the usable speech extracted from a co-channel utterance is less than the length of the utterance, which also depends upon the severity of the co-channel corruption. For this reason, it was necessary to determine the performance of the SID system under the condition of reduced speech availability. Shown in *Figure 2* are the results of an experiment designed to determine the SID performance with varying lengths of available speech. This experiment involves incrementally decreasing test data by truncating voiced-only test files. Several different scenarios were tested.

The situation where the SID system was trained with normal speech and tested with normal speech (indicated by ‘normal-normal’ *Figure 2*) is included as a control set. The other situations include testing with voiced-only segments when training both on normal speech and training on voiced-only speech.

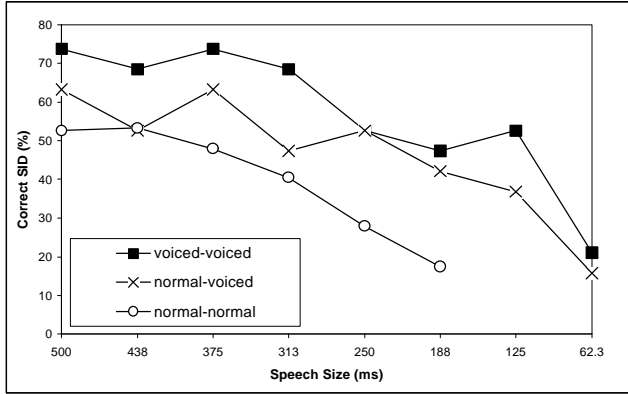


Figure 2: Speaker ID Accuracy vs. Varying Test File Size

It is interesting to observe that training and testing on voiced speech results in an overall higher SID accuracy than training and testing on normal speech. This is a clear indication of the high information content of the voiced only files. For the voiced-voiced situation, acceptable performance is attained for 300 ms or more of usable speech segments. It is promising to note that SID accuracy is still above 50% when only 125 ms (approximately 4 frames) of usable speech data is available.

### TIR Extracted Usable Speech

For co-channel speech segments to be considered usable, TIR has been selected at 20 dB across the entire utterance [1]. However, this is unlike the current work where the overall TIR was set to 0 dB and the frame-by-frame extraction is based on an incremented TIR threshold. TIR thresholds are selected as follows: 0 dB, 10 dB, 20 dB, and 30 dB. Co-channel speech is created by combining 5 concatenated speech files for each of 38 speakers from the TIMIT database at an overall TIR of 0 dB (equal overall power for both speakers). The 38 speakers are separated into two groups of 19 speakers each. Group A contains 14 female speakers and 5 male speakers. Group B contains 19 male speakers. In order to create the co-channel speech, the speakers from the groups are then combined according to the following criteria, where the letter ‘i’ indicates the speaker number within each group:

1. Group A(i) + Group A(i+1) to create 19 co-channel utterances at 0 dB overall TIR
2. Group B(i) + Group B(i+1) to create 19 co-channel utterances at 0 dB overall TIR

3. Group A(i) + Group B(i) to create 19 co-channel utterances at 0 dB overall TIR

As the frame TIR increases, the extracted speech is “cleaner,” but fewer frames meet the more stringent TIR requirement. Therefore, as the TIR threshold increases, the amount of usable speech extracted decreases. An examination of *Table 2* shows that there is about 5 times more speech available to the SID system in the 0 dB TIR case than the 30 dB TIR case.

Table 2: Amount of Extracted Usable Speech vs. TIR Threshold

| Frame TIR (dB) | unconstrained mean file size (seconds) | constrained mean file size (seconds) |
|----------------|--|--------------------------------------|
| 0              | 7.23                                   | 1.39                                 |
| 10             | 4.34                                   | 1.39                                 |
| 20             | 2.51                                   | 1.39                                 |
| 30             | 1.39                                   | 1.39                                 |

Because it is known from *Figure 2* that SID is dependent upon the amount of speech available, a second data set has been added where the mean file size for all TIR threshold increments was constrained. This way, the effects of varying the TIR can be investigated independently of the amount of usable speech available. It is interesting to note that SID accuracy increases as the frame TIR threshold increases, for the constrained file size situation (*Figure 3*). Speech frames extracted at higher frame TIR thresholds represent frames of “cleaner” speech; therefore, we expect the SID to perform better as TIR increases. It can be concluded with confidence that increasing the TIR improves SID accuracy for a fixed amount of usable speech.

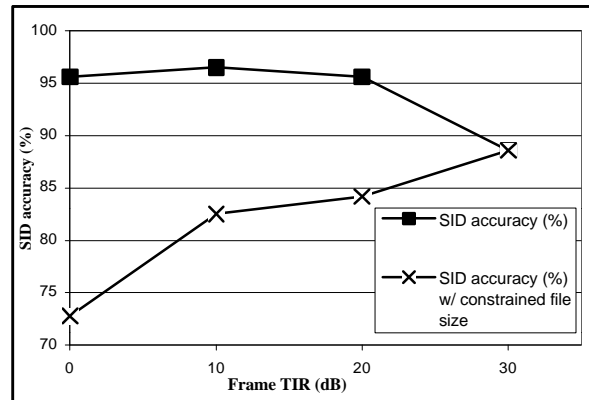


Figure 3: Speaker ID Accuracy with TIR Extracted Usable Speech vs. Frame TIR in dB

### SAR Extracted Usable Speech

Recently, the SAR method for detecting usable speech frames in co-channel speech was introduced. [3]. In order to determine the usability of frames

extracted using the SAR method, usable segments are extracted at SAR threshold values of 2, 5, 10, and 15. Usable segments are then assigned to either speaker based on a TIR of 0 dB.

Figure 4 shows that as the SAR threshold increases, there is a general decrease in the SID accuracy for both the constrained and unconstrained file size cases. It can be concluded that as the SAR threshold increases, fewer frames meet the SAR limitation, displayed by the unconstrained mean file size in Table 3. Therefore, a lower SAR threshold is desired because this allows for better accuracy. The SID accuracy should approach that of the constrained 0 dB TIR case as the SAR threshold approaches 0, because the usable segments are first selected using the SAR method, then assigned to either the target or interfering speaker using the 0 dB TIR method.

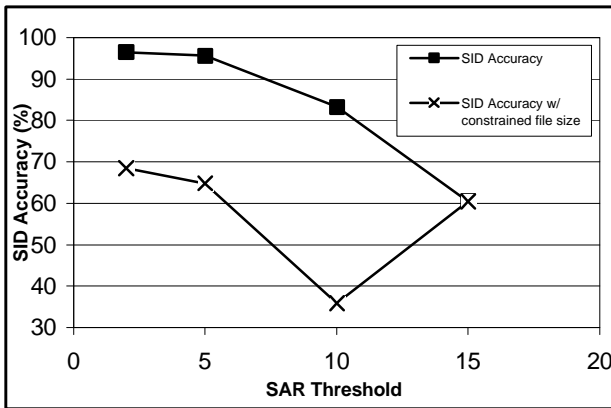


Figure 4: Speaker ID Accuracy with SAR Extracted Usable Speech vs. SAR Threshold

Table 3: Amount of Extracted Usable Speech vs. SAR Threshold

| Frame SAR (dB) | unconstrained mean file size (seconds) | constrained mean file size (seconds) |
|----------------|--|--------------------------------------|
| 2              | 6.42                                   | 1.3                                  |
| 5              | 5.41                                   | 1.3                                  |
| 10             | 2.11                                   | 1.3                                  |
| 15             | 1.31                                   | 1.3                                  |

#### 4. SUMMARY

This paper addresses the issues of applying the usable speech concept in the context of automatic speaker identification. The data presented here shows that voiced speech contains most of the useful information for SID, and that even small amounts of voiced data provide adequate SID accuracy. In addition, higher TIR thresholds provide ‘cleaner’ but fewer frames of usable speech. Finally, the ability of the SAR method to determine usable frames is shown to decrease as the SAR threshold increases.

#### ACKNOWLEDGEMENT

Effort sponsored by the Air Force Research Laboratory, Air Force Material Command, 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 representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory, or the U.S. Government.

#### 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] Yantorno, R. E., "Co-channel speech study", Final report for Summer Research Faculty Program, Research Laboratory AFRL/IF, Speech Processing Lab, Rome Labs, New York, 1999.
- [3] 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", ICSPACS, 2000.
- [4] Morgan, D.P.; George, E.B.; Lee, L.T.; Kay, S.M., "Cochannel Speaker Separation by Harmonic Enhancement and Suppression", Speech and Audio Processing, IEEE Transactions on , Volume: 5 Issue: 5 , Sept. 1997 Page(s): 407 - 424
- [5] Ma, K.W.; Zavaliagkos, G.; Meteer, M. "Sub-sentence discourse models for conversational speech recognition", Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on , Volume: 2 , 1998
- [6] Yantorno, R. E. Krishnamachari K. R., Lovekin J. M., Benincasa, D. B., Wenndt, S. J. "The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A Usable Speech Measure Employed as a Co-Channel Detection System", IEEE Intelligent Signal Processing 2000 [submitted]