

# Speaker Identification Enhancement under Co-channel Conditions using Sinusoidal Model and ESPRIT Harmonicity based Usable Speech Detection

Saurabh S. Khanwalkar, Brett Y. Smolenski and Robert E. Yantorno

Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, PA 19122-6077, USA

Email: [saurabhk@temple.edu](mailto:saurabhk@temple.edu), [bsmolens@temple.edu](mailto:bsmolens@temple.edu), [robert.yantorno@temple.edu](mailto:robert.yantorno@temple.edu)  
[http://www.temple.edu/speech\\_lab](http://www.temple.edu/speech_lab)

**Abstract:** The accuracy of present day speaker identification systems (SID) is degraded in the adverse acoustical environments i.e., by different kinds of interferences. The idea of usable speech is to identify and extract those portions of degraded speech which are considered useful for speaker identification. Recently, a usable speech extraction system was proposed to classify co-channel speech as usable speech and unusable speech for speaker identification. Speech segments can be declared usable based upon a Target-to-Interferer energy ratio (TIR). Studies indicate that a significant amount of the co-channel data is considered usable for speaker identification. By considering only usable speech for speaker identification instead of the corrupted co-channel speech, it is seen that there is an increase in the accuracy of speaker identification. A novel usable speech detection measure using the *sinusoidal model of speech* and ESPRIT (*Estimation of Signal Parameters via Rotational Invariance Technique*) Spectral Estimation is proposed and investigated, which resulted in 82% correct detection of usable speech segments based on TIR. The usable speech frames extracted using ESPRIT when tested with the speaker identification system, resulted in 84% accuracy in detecting speaker identity as compared to using entire co-channel speech which resulted in only 45% accuracy.

## 1. Introduction

Speaker identification plays an important role in electronic authentication. In an operational environment speech is degraded by many kinds of interferences. The interference can be classified broadly as stationary or non-stationary. Stationary interference is basically noise which can be dealt with by using de-noising and noise reduction techniques; whereas non-stationary interference can be speech data from a different speaker or non-stationary noise. Such interference is a common occurrence and the corrupted speech is known as co-channel speech [1]. Traditional methods of co-channel speech processing have been to enhance the prominent speaker (target), suppress the interfering speaker speech or both. However, previous studies on co-channel speech have shown that it is desirable to process only portions of the co-channel speech which are minimally degraded [2]. Such portions of speech considered usable for speaker identification are referred to as “usable speech”.

The knowledge of acoustical speech features in voiced or unvoiced speech plays an important role in many speech analysis-synthesis and speaker identification systems. A significant amount of research has been conducted in

finding reliable and accurate voicing determination in the past. Despite the numerous approaches that have been proposed to address this problem, it still remains an active research area because of the difficulties dealing with non-stationary properties of speech signal. A parametric signal model is often used to decompose a signal into a form that is more easily or efficiently processed than the original. As an example, “pitch-excited” *linear predictive coding* (LPC) models speech as the output of a linear, time-varying filter excited by an excitation pulse train (voiced speech) or a random noise source (unvoiced speech). This is a prime example of a model which is sufficient to accurately represent a signal, yet still decomposes the information involved into a reduced form which is easier to interpret and modify. However, the implicit assumptions made about the excitation signal in this model are quite restrictive, and result in speech which may sound “unnatural” for input signals which do not exactly match their assumptions. The desire for an alternative speech model which is both an “efficient” and a more general representation led to the development of the *sinusoidal model* of speech.

Many sounds of importance to human listeners have a pseudo-periodic structure, that is, over certain stretches of time, the waveform is a slightly-modified copy of what it was some fixed time earlier, where this fixed time period is typically in the range of 10 - 0.2ms, which corresponds to a fundamental frequency of 100 Hz - 5 kHz. Periodic signals can be approximated by a sum of sinusoids whose frequencies are integer multiples of the fundamental frequency. This gave rise to the concept of sinusoidal modeling of speech signals as introduced by McAulay and Quatieri [3], [4]. The key idea behind sinusoidal modeling is to represent the sinusoidal Fourier components or harmonics that are an equivalent representation of the sound waveform, explicitly and separately as a set of frequency and magnitude values. In general, the methods to estimate the spectral content of a signal are either parametric or non-parametric. Parametric methods are those which take advantage of known parameters of the signal, whereas the non-parametric methods make no such assumptions. The non-parametric methods used for frequency estimation of a sinusoidal signal provide good spectral resolution, but break down quickly in the presence of noise. This paper presents a parametric sinusoidal model based approach, which detects sinusoids in noise using the (Total Least Squares) TLS - ESPRIT Algorithm [5].

Speaker identification with usable speech segments has many associated problems. The usable extraction system cannot be a 100% accurate system as it relies on statistical approaches. Several measures have been developed to identify usable segments from degraded speech [6], [7], [8], [9], [10] and [11]. However, it is difficult to develop a classifier with a high identification performance rate due to the fact that the speech is diverse, and therefore, one is confronted with much different energy, transition regions, and sounds which may not have steady state regions. This means that all the extracted frames are not usable frames. The unusable frames extracted reduce the effectiveness of the speaker identification system. To overcome this problem there could be two possible solutions 1) enhance the extracted data and 2) develop more robust speaker identification techniques. A new approach to developing a usable speech measure is to use sinusoidal model-based approach for classification of speech into usable and unusable segments for Speaker Identification system [12].

## 2. Background

### 2.1. Sinusoidal Model of Speech Signal

Using the very well known signal analysis method of Fourier Series, the speech signal  $x(n)$  is modeled as the sum of a small number of sinusoids with time-varying amplitudes and frequencies in the presence of noise  $z(n)$ .

$$x(n) = \sum_{i=1}^p A_i e^{j(2\pi f_i n + \phi_i)} + z(n) \dots \dots \dots (1)$$

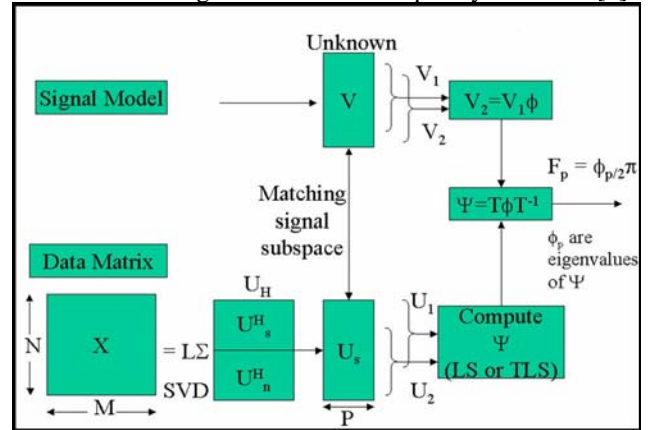
**Equation (1)** above is the basic sinewave model that can be thought of as speech-independent, i.e., the model can be applied to any signal [3]. The main step in sinusoidal modeling of speech is to develop a robust procedure for extracting the amplitudes and frequencies of the component sinewaves from the speech waveform. To find these parameters of the model, the short-time Fourier transform (STFT) (i.e. the DFT of a windowed frame of the signal) is calculated, and the peaks of the spectral magnitude are selected. In the terms of voicing determination, the voiced speech is periodic in structure and the unvoiced speech is noise-like in structure. When we consider the sinusoidal model of speech, the degree to which a given frame of speech is voiced is determined by the degree to which the harmonic (sinusoidal) model fits the original sinewave data.

### 2.2. Harmonic ESPRIT Power Spectrum

One of the problems associated with STFT is the *windowing* function. Windowing introduces side-lobes, which can mask the weaker signals. Trying to reduce the side-lobes eventually leads to a wider main lobe which means a reduction in frequency resolution. Further, the resolution in time and frequency is limited by Heisenberg's uncertainty theorem, which states that the area of the rectangle defined by the windowing function has a minimum given by  $\Delta f * \Delta t \gg \frac{1}{2}$ . Thus windowing of data clearly limits the frequency resolution obtainable, and a choice arises between high frequency resolution and high

time resolution, but not both. This led to the use of parametric methods such as AR models and eigendecomposition for harmonic modeling where there is no need to window the data.

The ESPRIT algorithm is a signal-subspace based frequency estimation technique that is built upon the principle of eigendecomposition and it also exploits the principle of signal subspaces [5], [13], [14]. This method is based on the decomposition of a vector space of a noisy signal by applying the eigendecomposition to the correlation matrix. However, since the second-order statistics are estimated from a number of noisy vectors, a better approach is to organize the vectors in a data matrix (such as Toeplitz or Hankel) and perform Singular Value Decomposition (SVD) [14] which reduces the computational complexity. **Figure 1** shows the block diagram of the proposed ESPRIT algorithm using the SVD. **Figure 1** shows the link between the signal model (i.e. the sinusoidal model of speech) and the data matrix (i.e. the  $N \times M$  block model of the same speech signal). It demonstrates the flow of the algorithm starting from the data matrix through the harmonic frequency estimates [8].



**Figure 1:** SVD and TLS ESPRIT algorithm to obtain the power spectrum of a particular frame of speech.

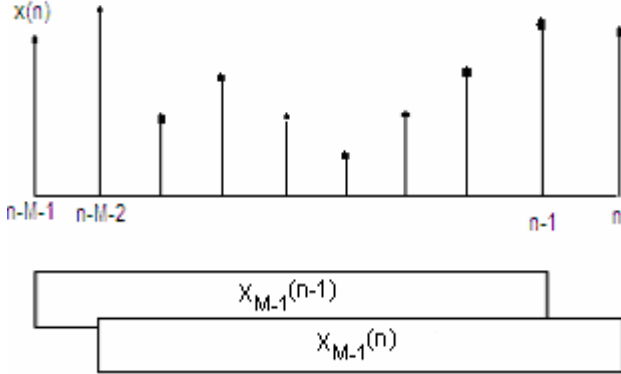
The essence of ESPRIT spectral estimation lies in the rotational property between the staggered subspaces (i.e. spread out over a period of time), that is invoked to produce the frequency estimates. In the case of the discrete-time signal or the vector model of the signal, this property relies on the observations of the signal over two identical intervals staggered in time [14].

With respect to the sinusoidal model of speech, if we consider a single complex exponential  $s_0(n) = e^{j2\pi f n}$  with complex amplitude  $\alpha$  and frequency  $f$ , this signal has the following property

$$s_0(n+1) = \alpha e^{j2\pi f(n+1)} = s_0(n) e^{j2\pi f} \dots \dots \dots (2)$$

that is, the next sample value is a phase-shifted version of the current value. This phase shift is represented as a rotation on the unit circle  $e^{j2\pi f}$ . In **Figure 2**, the block of signal vector model  $V$  is split into 2 vectors  $V_1$  and  $V_2$  with matrix  $\phi$  being the diagonal matrix of phase shifts between neighboring time samples of the individual,

complex exponential components of  $s(n)$ . Since the  $p$  columns of vector  $V$  are the length- $M$  time-window frequency vectors of the complex exponentials (sinusoids), i.e.  $V = [v(f_1) v(f_2) \dots v(f_p)]$ , we obtain the matrix  $\phi = \text{diag} \{ \phi_1, \phi_2, \dots, \phi_p \}$  where  $\phi_p = e^{j2\pi f_p}$  for  $p = 1, 2, \dots, P$ . Thus, since the frequencies of the complex sinusoids  $f_p$  completely describe this rotation matrix, frequency estimates can be obtained by finding this rotation matrix  $\phi$ .

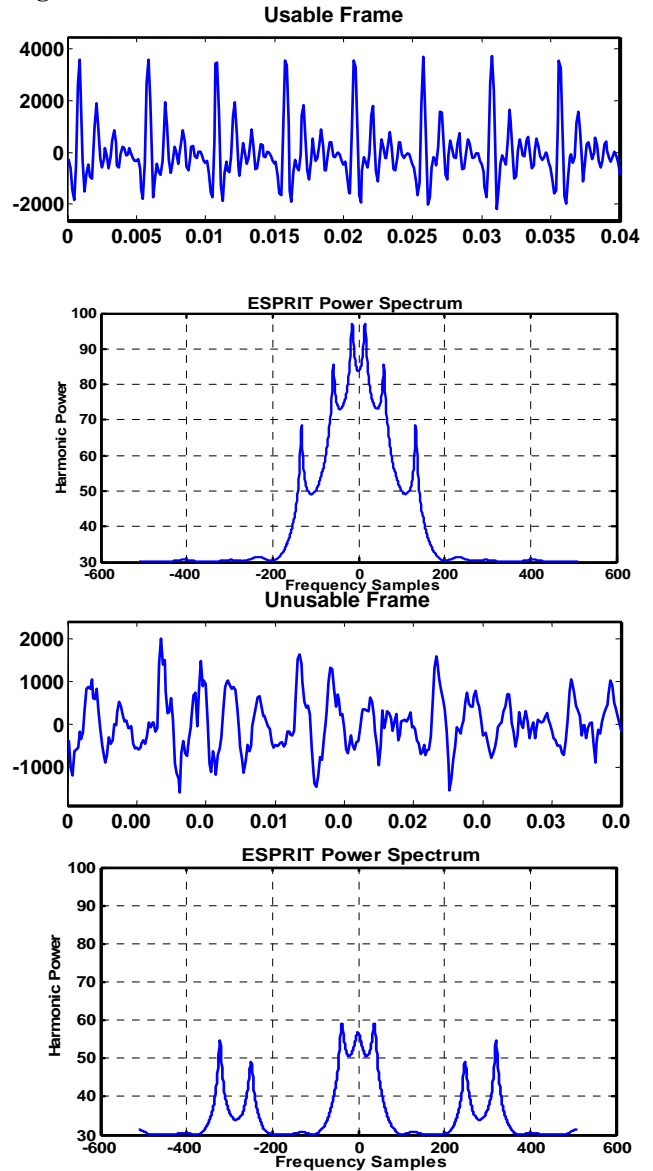


**Figure 2:** Time-staggered, overlapping windows used by the ESPRIT algorithm

ESPRIT algorithm makes use of singular value decomposition (SVD) for estimation of the sinusoidal frequencies and the respective amplitudes. The TLS technique is used to minimize the sum of squares of the error in the frequency estimation. As shown in the **Fig 1**, the data matrix  $X$  is considered to be composed of a signal subspace and a corresponding noise subspace. Using SVD, these 2 subspaces are separated as shown into  $U_s^H$  and  $U_n^H$  from the total subspace of  $U^H$ . To exploit the rotational property,  $U_s^H$  is partitioned into staggered subspaces  $U_1$  and  $U_2$  using SVD a second time. The frequency estimates are then obtained by finding the roots of the singular values. This operation is carried out iteratively using total least squares for error minimization.

In the case of a speech, the information about the harmonics can be obtained from the ESPRIT short-time spectral envelope of the speech signal. Experiments were performed for the optimum short-time frame size for this block-based signal subspace technique which resulted in 160 samples (20 ms) per speech frame. The data matrix is then formed as shown in **Figure 1** having the dimensions of  $N \times M$ , where  $N = 141$  and  $M = 20$ . This creates a 20-dimension signal + noise subspace. The  $P$ -dimensional signal subspace is obtained by using the Minimum Description length (MDL) and Akaike Information Criterion (AIC) for estimating the number of sinusoids. MDL and AIC are order selection information theoretic methods based on the eigenvalues of the correlation matrix provided by Wax and Kailath [15]. AIC tends to overestimate the order of the model, whereas MDL is more consistent in order estimation. So, the minimum of AIC and MDL estimation is selected as the order of the model, i.e., the  $P$ -dimension signal subspace. The ESPRIT

harmonic power spectrum is then obtained as shown in **Figure 3**.



**Figure 3: The ESPRIT Power Spectrum:** The topmost panel shows the usable speech frame and 2<sup>nd</sup> panel shows the power spectrum for the upper usable speech frame; the 3<sup>rd</sup> panel shows the unusable speech frame and the bottom panel shows the power spectrum for the upper unusable speech frame.

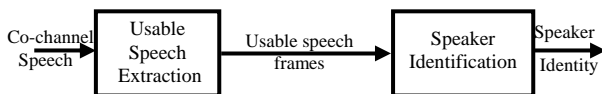
The frequency estimates of the  $P$  complex sinusoids are then taken as the  $P$  peaks in this pseudospectrum. Note the periodic structure of the speech data in the usable speech frame (top panel in **Figure 3**). The power spectrum obtained using the ESPRIT algorithm for this frame is shown in the 2nd panel. The power spectrum shows the harmonic peaks in the speech data of the usable speech frame. Note the lack of structure in the speech data of the unusable speech frame (third panel down in **Figure 3**). The harmonic power spectrum for this unusable speech frame is shown in the 4th panel. The power spectrum has spurious peaks

which do not exhibit any harmonicity. Note the harmonic power for the dominant peaks in both the cases. The usable data has relatively larger harmonic power as compared to unusable data.

Using an intelligent peak-picking algorithm, the dominant peaks from the ESPRIT spectrum are determined. Peak-picking analysis assumes that the spectral magnitude peaks will yield optimal sinusoidal component parameters, by degrading the sidelobe effects which generates spurious peaks. Peak-picking also limits the number of components to identifiable spectral peaks. The peak-picker was implemented based on a sorting method of selecting peaks. It first orders the peaks in decreasing order of harmonic power, and then selects the four highest peaks iteratively in the frame as the dominant peaks having the most harmonic content. These four peaks are then averaged to give the harmonic power content for that frame. The detection of usable speech segments is performed on frame-by-frame basis. The detection results are compared with those obtained using Target-to-Interferer Ratio (TIR) and the hits and false alarms are calculated.

### 3. Algorithm Flowchart

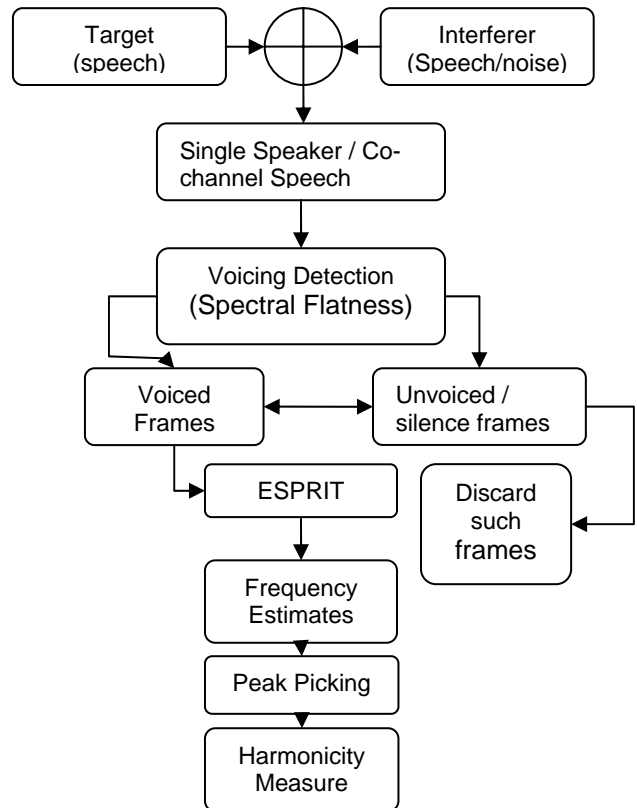
The block diagram of the usable speech extraction system and speaker identification is shown in **Figure 4**. The usable speech frames extracted are given as input the speaker identification system.



**Figure 4:** Block diagram showing application of Usable Speech detection

The various steps involved in obtaining this harmonicity measure are shown in the blocks shown in **Figure 5**. The usable speech segments are those portions of the speech which have the speaker's pitch information and hence have a periodic structure. Due to this nature of usable speech, only voiced speech is considered for the usable speech extraction. Spectral flatness method (SFM) is used as shown in **Figure 5** as a voiced-only speech extractor. The *Harmonicity Measure* which is the last block in the flowchart is where the Harmonic Power is computed.

The probability density function of the measure values for usable and unusable speech is computed to observe the amount of separability between the two. The threshold for the Usable Speech detection is determined based on this probability density function as discussed in the next section.

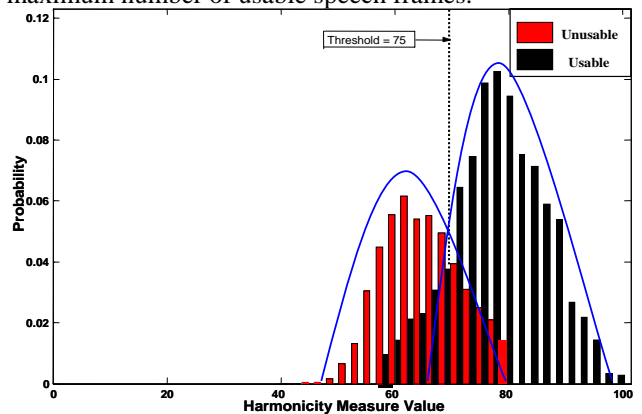


**Figure 5:** Flowchart of Steps in the Harmonicity Usable Speech Measure

### 4. Experimental Results

The ESPRIT harmonicity-based usable speech measure described above was tested on a closed set of 2145 co-channel utterances obtained from the TIMIT database. These utterances are created using 66 speech files with 33 male speech files and 33 female speech files.

To determine an optimum threshold for the proposed usable speech measure, the probability density function of the measure values is generated as shown in **Figure 6** below. The TIR threshold chosen is 15 dB as this gives the maximum number of usable speech frames.

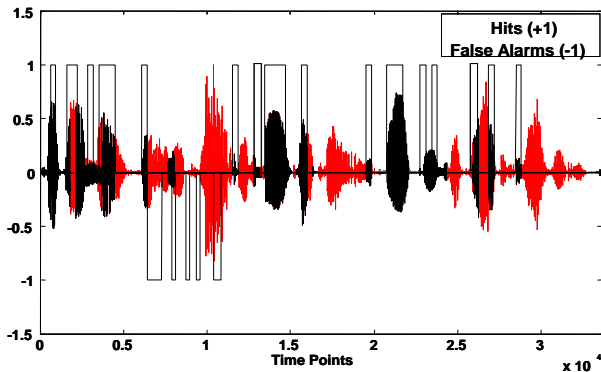


**Figure 6: Probability Density Distribution:** Average Harmonic Power in Each Speech Frame, Usable Speech (Black) and Unusable Speech (Red).

The black bars in the probability density function (**Figure 6**) represent occurrence of usable speech frames,

i.e., frames with  $TIR > |15dB|$ , and red bars represent the occurrences of unusable speech frames, i.e., frames with  $TIR < |15dB|$ . From the probability density function, a threshold value of 75 was found to be optimum.

For the determination of the effectiveness of the proposed measure, the percentage of hits and false alarms are calculated. The measure is said to have a hit, if the measure as well as the TIR value identify the frame as usable based on their respective thresholds. The measure is said to have a false alarm, if a frame is identified as usable by the measure, but unusable based on the TIR.



**Figure 7: Usable Speech Detection:** hits (shown by a value of +1) and false alarms (shown by a value of -1).

**Figure 7** shows the TIR identified usable speech frames (in black) and unusable speech frames (in red) of the co-channel data. In the plot, a hit is shown as a value of +1 and a false alarm is shown by -1.

**Table 1:** Comparison of Hits and False Alarms of the Previously Developed Usable Speech Measures with Esprit Usable Speech Measure.

Results	APPC	SAPVR	ESPRIT
Hits	74 %	71 %	82 %
False Alarms	26%	29 %	28%

**Table 1** above shows the comparison of the proposed usable speech measure with two previously developed usable speech measures. The proposed harmonicity usable speech measure resulted in 82% correct decisions or hits and 28% incorrect decisions or false alarms. The results are a significant improvement over the previous usable speech measures.

## 5. Speaker Identification Improvement under Co-Channel Conditions

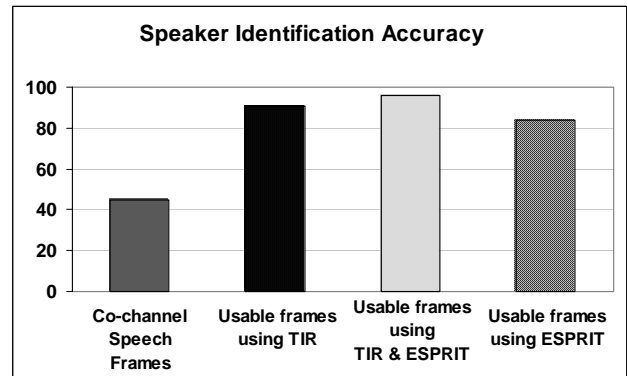
The objective of the research presented in this section is to study the performance of the speaker identification system with extracted usable speech segments from co-channel speech. The experiments were designed to emulate the speaker identification system using co-channel speech [16].

The frames extracted by the above algorithm were tested on the speaker identification system to evaluate the approach as a criterion for usability. Testing data was created from the extracted frames and used to drive the

speaker identification system. Frames were extracted in three ways and compared with entire co-channel frames.

- (1) Frames extracted with 15 dB TIR threshold,
- (2) Frames extracted with 15 dB TIR threshold with frames extracted by the proposed algorithm,
- (3) Frames extracted by the algorithm alone.

**Figure 8** shows the bar graph comparing the Speaker Identification accuracy with frames extracted using different criteria.



**Figure 8: Speaker Identification Accuracy:** Usable frames extracted using different criteria

It can be noted from **Figure 8** that the speaker identification accuracy has an increase when frames with good structure are added to TIR extracted frames. The correct speaker identification obtained using co-channel speech data alone is around 45%. Whereas, when tested with usable speech frames using 15dB TIR threshold, the speaker identification system resulted in 91% accuracy; and when tested with usable speech frames obtained from ESPRIT algorithm, there was 84% accuracy in speaker identification. When a combination of usable speech frames extracted from ESPRIT and TIR was used, a increased speaker identification accuracy of 96% was obtained. These experiments indicated that there is a definite increase in the accuracy of speaker identification system with the use of only usable speech instead of entire co-channel speech.

## 6. Discussion

The purpose of this paper was to identify the usable portions of co-channel speech in the context of speaker identification enhancement. It was found that the sinusoidal model of speech can be used as an effective method to identify usable speech segments. The proposed usable speech measure detected 82% of usable speech portions from co-channel speech data.

ESPRIT algorithm is based on the theory that the data matrix is made up of signal and noise subspaces. Further experimentation is needed be performed with speech data at lower SNR. Although, the proposed measure has a considerable percentage of false alarms; further improvement can be obtained using a higher order for the sinusoidal model, which will improve the spectral resolution, but at the expense of an increase in the computational complexity. As a next step in this research, the frequency estimates computed can also be used as a

measure of harmonicity in fusion with the harmonic power for further improvement in the accuracy of usable speech detection.

### Acknowledgements

The Air Force Research Laboratory, Air Force Material Command, and USAF sponsored this effort, under agreement number F30602-02-2-0501. 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.

### 11. References

- [1] R. E. Yantorno, Co-channel speech study, final report for summer research faculty program," tech. rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1999.
- [2] J. Lovekin, R. E. Yantorno, S. Benincasa, S. Wenndt and M. Huggins, "Developing usable speech criteria for speaker identification," *Proc. ICASSP* 2001, pp. 421-424, 2001.
- [3] McAulay R. J. and Quatieri T. F., "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-34, No. 4, pp. 744-754, August 1986.
- [4] McAulay R. J. and Quatieri T. F., "Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model," *IEEE Proc. Int. Conf. Acoustics, Speech, and Signal Proc.*, Albuquerque, NM, vol. 2, pp. 249-252, April 1990.
- [5] Roy, R., Paulraj, A., and Kailath, T. (1986). "ESPRIT: A Subspace Rotation Approach to Estimation of Parameters of Sinusoids in Noise," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-34, pp. 1340-1342, October
- [6] K. R. Krishnamachari and R. E. Yantorno, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions." *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Nov 2000.
- [7] J. M. Lovekin, K. R. Krishnamachari, and R. E. Yantorno, "Adjacent pitch period comparison (appe) as a usability measure of speech segments under co-channel conditions," *ISPACS* 2001.
- [8] N. Chandra and R. E. Yantorno, "Usable speech detection using modified spectral autocorrelation peak to valley ratio using the lpc residual," *4<sup>th</sup> IASTED International Conference Signal and Image Processing*, pp. 146{150, 2002.
- [9] A. R. Kizhanatham, R. E. Yantorno, and B. Y. Smolenski, "Peak difference autocorrelation of wavelet transform (pdawt) algorithm based usable speech measure," *IIS Systemic, Cybernetics and Informatics*, 2003.
- [10] N. Sundaram, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, "Usable speech detection using linear predictive analysis - a model-based approach," *IEEE International Symposium on Intelligent Signal Processing and Comm. Systems, ISPACS 2003*.
- [11] Iyer, A.N., Gleiter, M., Smolenski, B.Y., and Yantorno, R.E., "Structural Usable Speech Measures Using LPC Residual", *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Dec 2003
- [12] Yang Shao and DeLiang Wang (2003), "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," *ICASSP2003*, Volume 2, Pages 205-208
- [13] Y. Ephraim and H. L. Van Tress, "A Signal Subspace Approach for Speech Enhancement" in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Minneapolis, MN, pp. 355 - 358, April. 1993.
- [14] Manolakis G. D., Ingle K. V. and Kogan M. S., *Statistical and Adaptive Signal Processing: Spectral Estimation*, McGraw-Hill Science/Engineering/Math (December 1999)
- [15] Wax, M., and Kailath, T. (1985). "Detection of Signals by Information Theoretic Criteria," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP - 32, pp. 387 - 392, April.
- [16] A. N. Iyer, B.Y Smolenski, R. E. Yantorno, J. Cupples, S. Wenndt, "Speaker Identification Improvement Using The Usable Speech Concept," *European Signal Processing Conference (EUSIPCO 2004)* (Accepted).