Babble Noise: Modeling, Analysis, and Applications

Nitish Krishnamurthy, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Speech babble is one of the most challenging noise interference for all speech systems. Here, a systematic approach to model its underlying structure is proposed to further the existing knowledge of speech processing in noisy environments. This paper establishes a working foundation for the analysis and modeling of babble speech. We first address the underlying model for multiple speaker babble speech—considering the number of conversations versus the number of speakers contributing to babble. Next, based on this model, we develop an algorithm to detect the range of the number of speakers within an unknown babble speech sequence. Evaluation is performed using 110 h of data from the Switchboard corpus. The number of simultaneous conversations ranges from one to nine, or one to 18 subjects speaking. A speaker conversation stream detection rate in excess of 80% is achieved with a speaker window size of ± 1 speakers. Finally, the problem of in-set/out-of-set speaker recognition is considered in the context of interfering babble speech noise. Results are shown for test durations from 2–8 s, with babble speaker groups ranging from two to nine subjects. It is shown that by choosing the correct number of speakers in the background babble an overall average performance gain of 6.44% equal error rate can be obtained. This study represents effectively the first effort in developing an overall model for speech babble, and with this, contributions are made for speech system robustness in noise.**

***Index Terms*—Babble, multispeaker babble, noise analysis, noise characterization, speech analysis.**

I. Introduction

THERE has been significant research in the past to ensure speech system reliability in adverse conditions. Extensive work has been performed on robustness for automatic speech recognition (ASR) [1], speaker-ID [2], and other speech domains. Most studies consider robustness across variations in noise, stress, accent, dialect, and emotion. Hansen *et al.* [3] and Varadarajan *et al.* [4] developed algorithms for speech recognition and speaker identification systems that are robust to speech in noise (i.e., Lombard effect), stress and emotion. One of the most important aspects for system reliability is robustness to environmental noise. Approaches that utilize separate models for noise are explored in Akbacak and Hansen [5] and Varga and Moore [6]. These models use environment specific characteristics to provide robust speech system performance. One of the most challenging noise conditions is multispeaker or babble noise environment, where the interference is speech from speakers in the vicinity. This noise is uniquely challenging because of its highly time evolving structure and its similarity to the desired target speech. These difficulties have been well documented in many studies for robustness. Cooke [7] modeled consonant perception in babble with varying number of speakers. Li and Lutman [8] model the change in kurtosis as a function of speakers in babble. These studies have explored the nature of babble and its characterization for improved speech recognition and the impact of the number of speakers on speech recognition. The primary focus of this study is to develop a foundation to address babble, and in particular, a framework is proposed to detect the number of speakers in babble. Also, there have been a number of studies on the perception of multi-speaker babble in the field of audiology and hearing sciences. Loizou in [9, Ch. 4] describes the perceptual aspects of multispeaker babble, where it is noted that as the number of speakers increase, the ability to recognize monosyllables from individual speakers increase. Listeners exploit gaps or dips present in speech to recognize speech corrupted by multispeaker babble. For a lower background speaker count, speech recognition is better since there are more gaps in the speech voices. As the number of speakers increase, there are fewer gaps in the spectrogram making identification of individual speech difficult. For robust speech recognition in babble, Morales *et al.* [10] notes the special properties of speech masked with babble. There, it is shown that it is possible to improve speech recognition performance in unknown noisy conditions by masking the noise with known noise types. Of special interest was the competing speaker case that provided better performance when the competing speaker was masked with babble, suppressing the competing speaker in babble. Previous studies have shown how human and machines behave differently under different babble scenarios.

In the present study, we propose to formalize a framework to analyze babble (Section II). The focus here is not on competing speaker separation [11], where information from multiple microphones is used to separate speech from individual speakers, but on characterizing babble as observed using a single microphone. Here, the differences between real babble (e.g., collecting babble in real-life settings) and synthetic babble (e.g., adding separate single speaker speech files together) are considered. Next, we study the impact of multispeaker babble on in-set/out-of-set speaker identification systems. It is demonstrated that as expected, by choosing matched test-train conditions, better performance is obtained than with mismatched test-train conditions. Knowledge estimated from

Manuscript received May 28, 2008; revised December 25, 2008. Current version published July 31, 2009. This work was supported by the Air Force Research Laboratory under a subcontract to RADC, Inc., Under Grant FA8750-05-C-0029. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

The authors are with the Department of Electrical Engineering, Erik Johnson School of Engineering and Computer Science, Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2015084

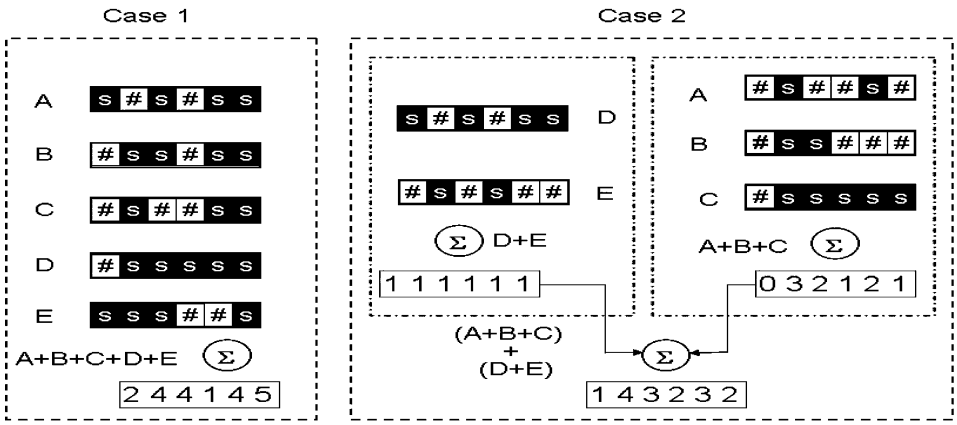


Fig. 1. Babble Noise: the left block shows five streams (five speakers) are overlapped and the right block shows the case where two conversations (two speakers and three speakers) are overlapped.

babble noise is shown to improve speech system performance under babble noise conditions.

II. Analysis/Modeling of Babble

The common notion of babble in speech systems is the noise encountered when a crowd or a group of people are talking together. An approximation to real babble is to add streams of speakers speaking individually, rather than adding conversations. There are some significant differences between such an approximation and real babble which, to our knowledge has not yet been considered. Consider a scenario with five speakers as shown in Fig. 1. Individual streams of five speech utterances are shown in Case 1 on the left, where s represents speech activity, and # silence. The speech frames are labeled 1 and the silence frames are labeled 0. If five speech streams are added, this implies all five subjects will speak simultaneously, but no two will be engaged in a conversation. In Case 2, it is assumed that in the room, the five are divided into two groups, one consisting of two subjects and the other of three subjects who are involved within two separate conversations. Over time, there will be babble from two conversation groups, where most of the time there would be simultaneous speech from two speakers, one from each conversation. Speakers involved in a conversation would change over time since they take turns to speak within each conversation. In Case 1, there would be five subjects talking simultaneously, whereas, in case 2 there would be two subjects talking simultaneously most of the time, and these two subjects would change with time, depending on the dynamics of each conversation. Fig. 2 shows the difference in the distributions (pdfs) of the number of speakers speaking per frame when two speakers are added versus two conversations are added. As observed from the pdfs, when speech from individual speakers are added there is no possibility of more than two speakers speaking at the same time whereas, when two conversations are added most of the time two speakers are speaking but at times it is possible that all four speakers speak simultaneously. So, to model babble noise, it is more accurate to employ a model consisting of a sum of conversations rather than individual speech streams of conversations overlapped with each other. When individual speech streams are overlapped under the assumption of independence, it is an inaccurate model for actual babble noise since speech

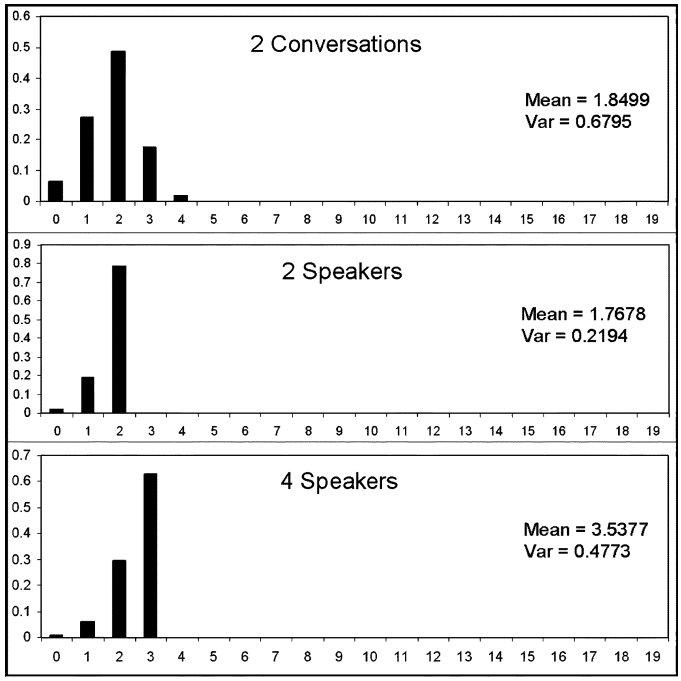


Fig. 2. Difference in pdfs of number of speakers talking simultaneously when (a) two conversations are added, and (b) two speakers speaking individually are added.

from each speaker in a conversation is correlated to the other (i.e., turn-taking within each conversation). The next section identifies the variables that influence babble in a given environment.

A. Factors Influencing Babble

Babble noise is a function of the number of speakers in an acoustic environment. The number of conversations and grouping of the speakers impact the acoustic variability of the babble. In a conversation, there can be more than two subjects participating, but usually there is only one subject speaking at any given point in time. In a conversation, the speaker might change with time, but in general there will be only one speaker speaking. The number of conversations dictate the number of subjects speaking simultaneously in babble. Reducing the

number of subjects speaking simultaneously will result in an increase in the acoustic variability. The number of conversations in the given environment influences the number of possible speakers speaking at any instant of time. In addition to the number of speakers, the emotion/stress levels in the individual conversations [12] play a role in the spectral structure of the individual data streams. The language of the individual conversations will also contribute to the structure of the individual conversations. The acoustics of the environment play a role in deciding if the individual sources contribute additively, or if there is a convolution/reverberation effect in the babble noise. The placement of the microphone relative to the individual conversations establishes the dominance of individual speakers in the recorded data. Another factor influencing babble noise in an environment is the timing/turn-taking nature of speakers within each conversation group. This will depend on the conversation topics and the number of individual speakers who contribute to each conversation. Within each conversation, the dominance of individual speakers will affect the nature of babble from a given environment. Given these factors, it can be deduced that the approximation of real babble data by adding individual sentences therefore depends upon the speech application and the specific kind of babble environment. Here, we focus on babble as a sum of conversations.

B. Analysis of Turn-Taking Within a Conversation

In this section, a model of babble as a sum of conversations is proposed. Here, the pdf distribution of the speech from a person A is $p(a)$, and it is assumed that speech streams are statistically independent and identically distributed. With this, the joint pdf of n streams is given by

$$\psi(a_1 + a_2 + a_3 \cdots + a_n) = \psi(\omega_a)^n \tag{1}$$

where ψ is the characteristic equation of the individual pdfs. Alternatively, if babble is modeled as a sum of n conversations, then the pdf of the speech stream output of the conversation is given by $\psi(a_1^1, a_1^2)$, This can be written as $\psi(\overrightarrow{a_1})$ since the speech from the speakers will be correlated. If babble is modeled as a sum of n conversations assuming conversations to be independent, then the joint characteristic equation of n conversations is given by

$$(\psi(\overrightarrow{\omega}))^n. \tag{2}$$

If each conversation is restricted to be between two people, the conversation output can be modeled as a sequence of $0's, 1's$, and $2's$. Here, 0 denotes silence, 1 denotes one subject talking, and 2 denotes both subjects talking. This decision is made on a frame-by-frame basis. Such a scheme lends itself to Markov modeling where each state is a conversation mode. In a conversation involving two subjects, it is expected that a single person talks most of the time, with silence between turn-taking and occasionally small instances where both speakers speak simultaneously. The situation where both produce speech simultaneously occurs when there is a short pause between turn taking and the frame overlaps at the end of one speaker and start of another. A separate case occurs when both are laughing, or agreeing, or if there is back-channel feedback, etc. If P_0, P_1 and P_2 are the probabilities of observing 0, 1, and 2 then intuitively, $P_1 > P_0 > P_2$. If we model babble as a sum of N conversations, then $2N+1$ states are possible (0 speakers to $2N$ subjects speaking per frame), the probabilities of each state individually is

$$\begin{aligned}
P_0^N &= P_0^N \times P_1^0 \times P_2^0 \\
P_1^N &= \binom{N}{N-1 \quad 1 \quad 0} P_0^{N-1} \times P_1^1 \times P_2^0 \\
P_2^N &= \binom{N}{N-2 \quad 2 \quad 0} P_0^{N-2} \times P_1^2 \times P_2^0 \\
&\quad + \binom{N}{N-1 \quad 0 \quad 1} P_0^{N-1} \times P_1^0 \times P_2^1 \\
&\vdots \\
P_N^N &= \binom{N}{0 \quad N \quad 0} P_0^0 \times P_1^N \times P_2^0 \\
&\vdots \\
P_N^2 N &= \binom{N}{0 \quad 0 \quad N} P_0^0 \times P_1^0 \times P_2^N
\end{aligned}$$

where

$$\binom{N}{a \quad b \quad c} = \frac{N!}{a!b!c!} \tag{3}$$

For a two-speaker case, it can be seen that unless $P_2 > 0.5 * P_1, P_2^2$ will be the most probable event when two streams are combined. This situation can be extended to N conversations where P_N is the most probable event. This observation is used to detect the number of speakers in babble conditions.

C. Analysis of Babble as a Function of Number of Speakers

For analysis of speech babble, babble is studied as three separate acoustic cases. Here, babble is categorized based on the number of speakers speaking instantaneously with the following three classes.

- Competing speaker (**COMPSPKR**): having only two subjects talking simultaneously.
- Babble (**BAB**): In this condition individual speakers can be heard and at times, individual words can also be heard.
- Large-crowd (**LCR**): Sounds like a diffused background rumble, where individual conversations or speakers are not distinguishable.

The boundaries between BAB and LCR are fluid and depending on various factors such as the relative distance of the conversations from the microphone, the category of babble noise can be decided. To obtain an estimate of the boundaries between BAB and LCR, a perceptual experiment was carried out. Here, each subject was given the definitions of BAB and LCR and asked to classify 18 samples as babble as BAB or LCR. In the sound samples, the number of speakers in babble was varied from two to ten. Three instances of sounds for each speaker count were generated. A total of 12 subjects were a part of the experiment. The results are shown in Table I. Each of the above mentioned babble scenarios have their unique features. In the babble scenario (BAB), individual speakers are generally not discernible but occasionally, individual words along with the speaker can

TABLE I
PERCEPTUAL CLASSIFICATION OF BABBLE (BAB) AND LARGE CROWD NOISE (LCR)

Number of Speakers in Babble	Babble	LargeCrowd
≤ 3	100%	0%
4	66%	34%
5	18%	82%
6	27%	73%
≥ 7	0	100%

be identified. The regions between four to six speakers in babble are the most confusable babble types; this is the transition region from babble to large crowd noise. As the number of speakers increase the probability of observing individual words reduces. In the large crowd scenario (LCR), individual speech or speaker information cannot be identified. In this case, LCR-Babble consists of speaker rumble where no specific information can be obtained (e.g., speaker count, conversation, individual words, etc.). As the number of subjects in babble increase, the time varying nature of the babble reduces. The change in properties of babble with an increase in the number of speakers is studied in the following sections.

III. BABBLE AND ACOUSTIC VOLUME

In the previous section, babble is modeled as audio streams from individual speakers. This section studies the impact of the overlap of phone sequences on the resulting acoustic space. As the number of overlapping phones increase within a given babble utterance, the differences between individual frames are averaged and their identity becomes blurred. This removes the ability to distinguish individual phones in a babble utterance. Fig. 3 demonstrates this aspect using the reduction in Itakura–Saito (IS) [13] distance. This distance reduces between waveforms when the number of distinct phones superimposed increases. The symmetric IS measure is defined as

$$d_{\mathrm{IS}(j)} = \frac{1}{2}\{d_j(\overrightarrow{a_s}, d_j(\hat{\overrightarrow{a_s}})) + d_j(\hat{\overrightarrow{a_s}}, \overrightarrow{a_s})\} \tag{4}$$

where $\overrightarrow{a_s}$ and $\hat{\overrightarrow{a_s}}$ are the all-pole model parameters from the gain normalized spectra of the two waveforms to be compared, and $d_j(\overrightarrow{a_s}, \hat{\overrightarrow{a_s}})$ is the IS distance given by

$$d_j(\overrightarrow{a_s}, \hat{\overrightarrow{a_s}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[e^{v(\omega)} - v(\omega) - 1\right] d\omega \tag{5}$$

where

$$v(\omega) = \log\left(\frac{\sigma_{\hat{\overrightarrow{a_s}}}}{|\, A(\hat{\overrightarrow{a_s}}, \omega)\,|^2}\right) - \log\left(\frac{\sigma_{(\overrightarrow{a_s})}}{|\, A(\overrightarrow{a_s}, \omega)\,|^2}\right). \tag{6}$$

The experiments are conducted using synthetic phones from the same speakers generated by the Festival Speech Synthesizer system. The phones generated are @, A, Y, U, i, where phones are represented using Single-Symbol ARPAbet version [14, pg. 117]. These phones are generated with 16-kHz sample rate for 12 ms. These waveforms are modeled using 12th-order linear prediction coefficients (LPCs) [15]. Fig. 3 illustrates the frequency response of the LP models as the number of overlapping phones is increased. Two observations can be made from this experiment: First, as the number of overlapping phones increase, the ability to distinguish between the phoneme spectra decreases implying that the resulting sounds are similar. This observation is also reflected in the IS measures between waveforms. Second, the resolution of resonating poles in the LP spectra are less distinct as the number of speakers in babble increase. As the number of speakers increase, the spectrum of babble converges to an aggregate model of the vocal tract configuration across different phones. Fig. 4 shows the mean IS distances and the variance in those distances as a function of number of overlapping phonemes. As the number of phonemes k increase, various combinations of five phones are chosen and superimposed. This process can be extrapolated to $K \rightarrow \infty$, with an infinite number of phonemes overlapping, the resulting spectra approximates speech shaped noise. There is a monotonic decrease in the mean and variance of the distances between the averaged phones as the number of phones in babble utterance increases. This suggests that with an increase in the number of speakers in babble, the noise becomes localized in the acoustic space. Here, the acoustic space is characterized by the LP coefficients. This observation can also be extended to general acoustic spaces. Let $X_1, \ldots\ldots, X_k$ be N-dimensional vectors describing the acoustic space of the given data. It is assumed that the centroids of the vector quantized acoustic features sufficiently describe the acoustic space. It is noted that most speech systems are based on some form of classification for which a prerequisite step is quantization of the available acoustic space. For any acoustic space, the farther the entities to be classified, the better is the classification accuracy. An N dimensional cube is used to model the acoustic space enclosed by these centroids. Fig. 5 describes the construction of this space in two dimensions. The vertices of this figure are given by $(x_{\max}, y_{\max}), (x_{\min}, y_{\max}), (x_{\min}, y_{\max})$, and $(x_{\max}, y_{\min})$. In this N-dimensional space, the hyper-cuboid would have 2^N vertices, where the cuboid space is totally characterized by the following set of points:

$$\{\arg\max x_1, \arg\min x_1, \ldots, \arg\max x_N, \arg\min x_N\}. \tag{7}$$

Here, the maxima and minima are evaluated for each dimension separately across all centroids. The entire acoustic space of the data is enclosed within a volume bounded by these extreme points. Since the space is modeled using a cuboid, all the centroids are either on the edges or within the volume enclosed by the cuboid. The volume of this cuboid is measured, and this volume will be an indicator of the acoustic variation of the data. The volume of this enclosed N-dimensional cubiod with adjacent edges $e_1, e_2, e_3, \ldots, e_N$ is given by

$$V = e_1 * e_2 * e_3 * \ldots\ldots * e_N \tag{8}$$

where

$$e_1 = \arg\max x_1 - \arg\min x_1. \tag{9}$$

Here, it is noted that a large acoustic volume implies an expansive acoustic variation in the data. Conversely, a small

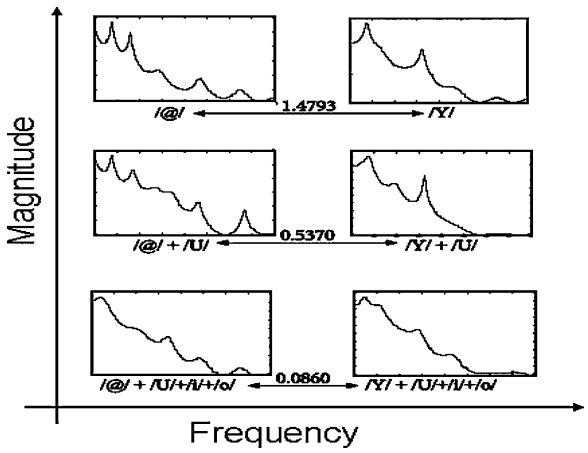


Fig. 3. IS measure decreases as the number of superimposing phones increase.

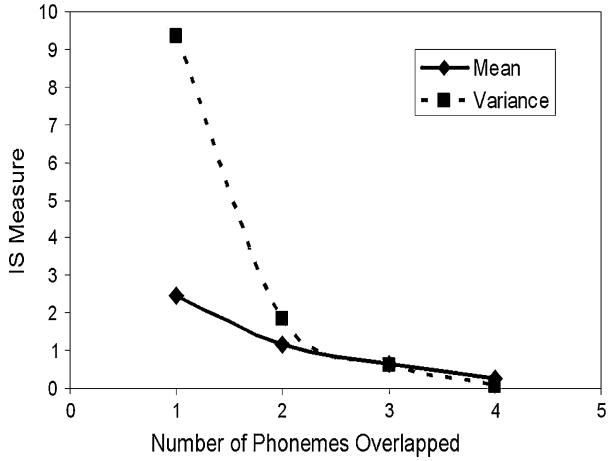


Fig. 4. As the number of superimposing phones increase, the mean spectral distance reduces. The individual spectra of superimposed phones are less distinguishable as seen from the drop in variance.

acoustic volume would mean less acoustic variation. For a single speaker, a larger acoustic space is expected since distinct phonemes would be present. Whereas, for babble with multiple simultaneous speakers, the expected acoustic volume should be smaller. Furthermore, as the number of speakers in the babble increase, a shrinkage in the acoustic space is expected. Another measure of this spread of the acoustic space is the distance between the pdf centroids. These centroids are an estimate of the compactness of the acoustic data clusters. This scheme is illustrated for one centroid in Fig. 6. The Euclidean distance between two points in the N-dimensional space is

$$d(x,y) = \left(\sum (x(i) - y(i)\right) (x(i) - y(i))^T)^{\frac{1}{2}} \qquad (10)$$

These distances are calculated for all centroids describing the acoustic space. As the number of speakers increase within babble, the centroid clusters will move closer (e.g., the points A, B, C, D, E in Fig. 5 will move closer together). This metric therefore provides additional information on the distribution of the centroids (i.e., information pertaining to relative closeness of the centroids in the acoustic space). These volume and acoustic space characteristics are evaluated on a synthetic babble corpus constructed using the test corpus of TIMIT consisting of both male and female speakers. The number of speakers is varied uniformly from one to nine subjects speaking at a time. Here, 19-dimensional Mel Frequency cepstral coefficients (MFCCs) are then extracted from 125-ms (1000 samples at 8-kHz sample rate) frames. The large frame size has been chosen to analyze the aggregate spectral structure of babble. These MFCC vectors are assumed to characterize the acoustic space of babble by clustering and employing Gaussian mixture models (GMMs). The pdfs are given as

$$p(x) = \sum_{j=1}^{m} \omega_j N(x \,|\, \mu_j, \sigma_j) \qquad (11)$$

where $(N(\,))$ is the conditional 19-dimensional Gaussian. The GMM model parameters are estimated using the EM algorithm, where the data is split into 32 mixtures[1] and the means of each mixture is used to characterize the acoustic space. The acoustic volume is evaluated using these centroids. Fig. 7 shows the resulting monotonic decrease in the acoustic volume as the number of speakers in babble increases. Here, there is an exponential reduction in volume as the number of speakers in babble increase. To process speech in noise, ideally, noise should be localized in this space and separated from the acoustic volume. However, noise and speech share the same acoustic space when described using MFCC spectral features, therefore distinguishing speech versus babble becomes difficult. Moreover, the acoustic space of babble is a subregion of the entire acoustic space occupied by speech from a single speaker. Fig. 8 shows the histograms of distances between the centroids for a speech signal with one, four, and nine speakers. The distance histograms with one speaker is more broad and flat, with distributions approximating Gamma distributions as the number of speakers increases. The variation of the mean distances is shown in Fig. 9, where as the number of speakers increase, the mean distance between the centroids decreases, which implies the acoustic features are clustered tightly. As is evident from the volume and distance plots, in cases where there is a reduced number of speakers in babble, the centroids enclose a larger volume, and they are uniformly distributed. With an increase in the number of speakers, the mean distance reduces and the volume also decreases. The acoustic volume describes the reduction in the acoustic space of babble with an increase in the number of contributing speakers. Also, another impact of the increase in the number of speakers is an increase in the abruptness in the spectral movement for babble which is studied in the next section.

IV. Acoustic Movement in Babble

As observed in the previous section, the amount of acoustic variability of babble depends on the number of subjects contributing to the babble. If speech from a subject is modeled as a sequence of phoneme utterances, multispeaker babble can be

[1]It is noted that 19 dimensions were used in [16], and with 32 Gaussian mixtures the likelihoods were found to converge.

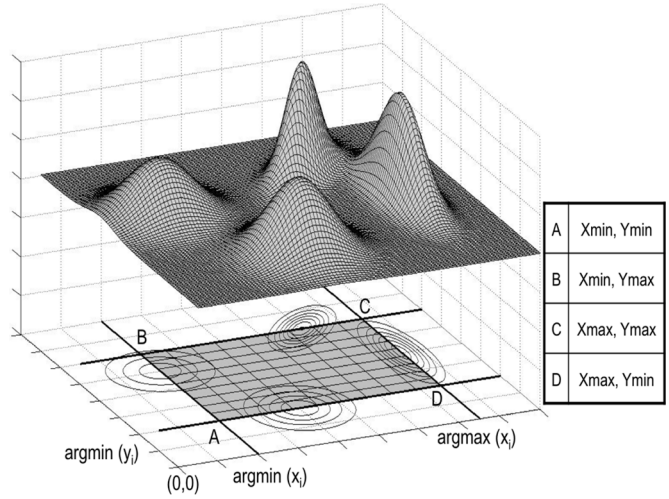


Fig. 5. Illustration of the acoustic area/volume occupied by a GMM of four mixtures.

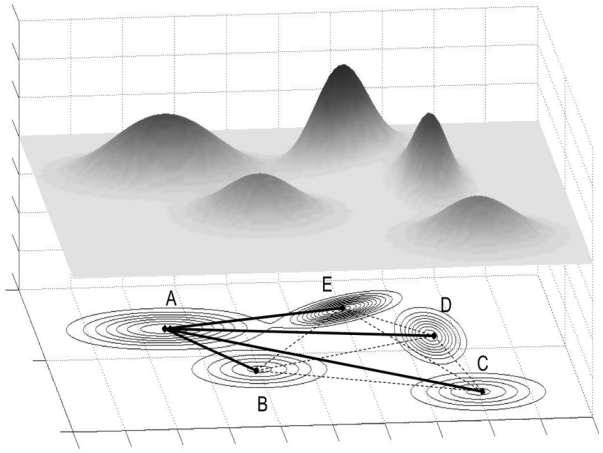


Fig. 6. Inter-centroidal distance between centroids of a four-mixture GMM.

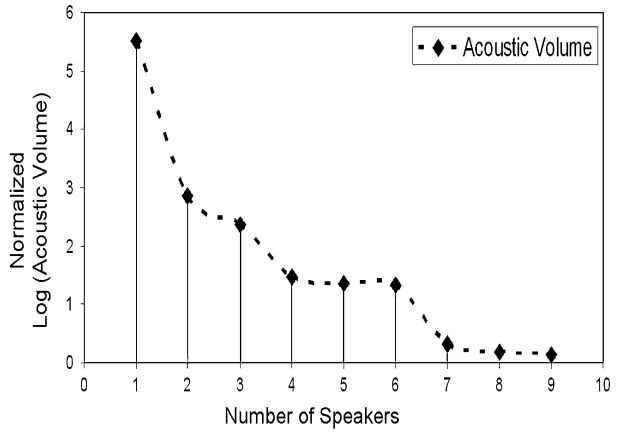


Fig. 7. As the number of participating speakers in babble increase, the volume enclosed by their GMM centroids reduces.

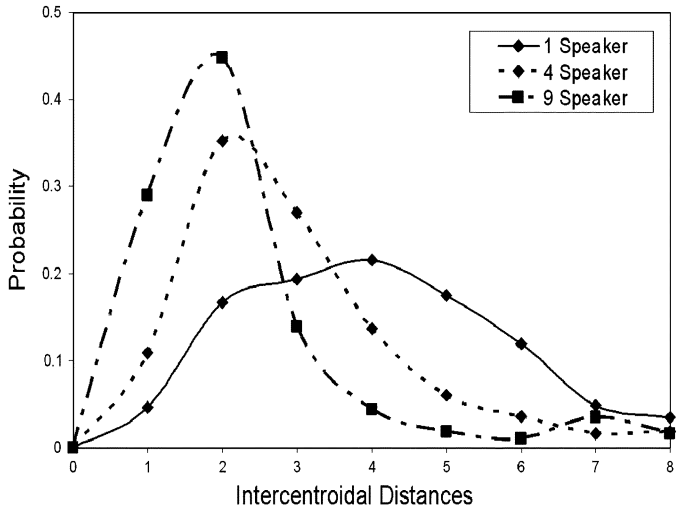


Fig. 8. Skewness in the pdfs of the inter-centroidal distance increases as the number of speakers in babble increase showing the nonuniform spread of data in acoustic space.

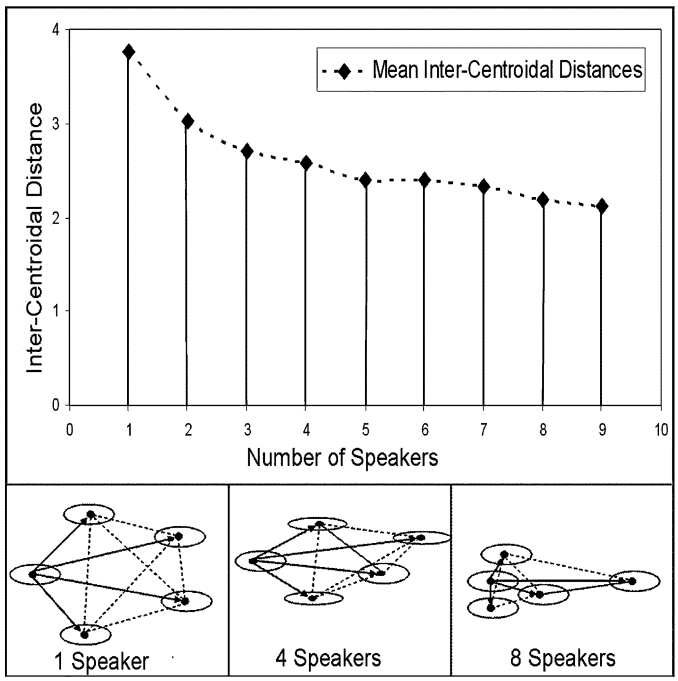


Fig. 9. Top: centroids cluster closer as the number of speakers in babble increase. Bottom: figure illustrating the resulting compactness of the acoustic space with an increase in the number of speakers in babble.

viewed as a layering of phonemes and silence periods from individual subjects. The acoustic trajectory of speech from a single subject is expected to be smooth for a majority of the portions since the inertial nature of the physiology of speech production would not allow for frequent abrupt movement in the acoustic space. Trajectory models of speech capitalize on this phenomenon. Gish [17] considered this acoustic trajectory as movement in the feature space (the trajectory is modeled as a polynomial to fit features in a window parametric trajectory modeling) and Gong [18] considered this as movement within the states of an HMM (this is done by assigning Viterbi paths within HMMs stochastic trajectory modeling). If we consider the acoustic trajectory of babble, abrupt and uneven trajectories are expected in contrast with natural speech from a single speaker. It is suggested that this is due to the layering of individual speech trajectories, resulting in conflicting articulatory responses from simultaneous speakers. A direct consequence of the trajectory being

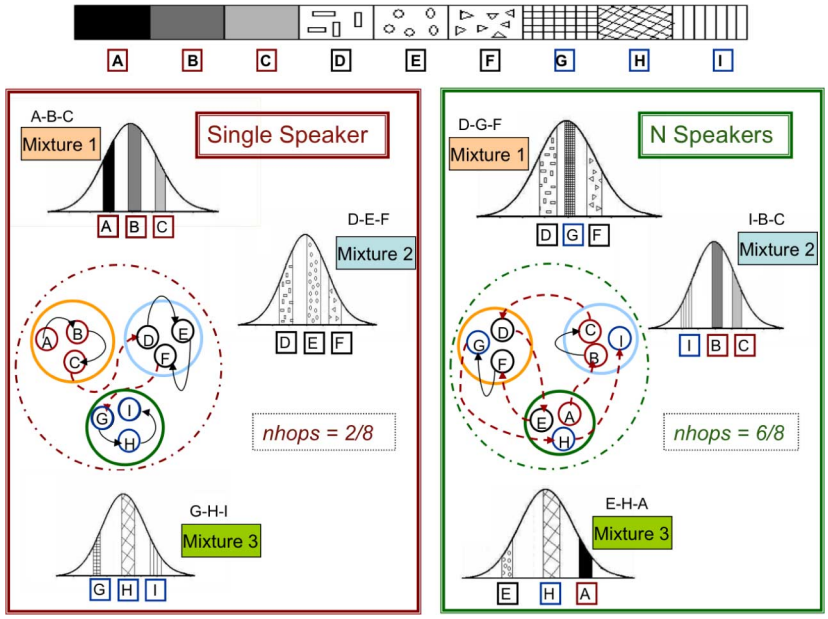


Fig. 10. Illustration of the reduction in the contiguity of adjacent frames as number of subjects in babble increase.

smooth is that individual features would be localized in the feature space. The abrupt or random nature of babble would lead to a relatively smaller localized acoustic space. The acoustic trajectory is a time function of the variation of acoustic features given by

$$f(i) = d(\ldots, x_{i+1}, x_i, x_{i-1}, \ldots\ldots\ldots) \tag{12}$$

where d is a function that maps the feature space to the trajectory space. In a quantized acoustic space, the features from the same acoustic region will share similar acoustic properties. The function d is defined as

$$d(i) = \begin{cases} 1, & \text{when } x_i = x_{i-1} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

and is an indicator of movement between quantization regions. Here, a "1" indicates movement across quantized regions and a "0" means that the current frame is in the same quantization region as the previous frame. Acoustic features are thus mapped into a sequence of zeros and ones, where a large number of 0's would signify blocks of contiguous speech from a consistent speaker, while a series of 1's suggests more random movement between speakers and phoneme content. Fig. 10 illustrates the acoustic movement for a single speaker and multispeaker babble. Here, A-B-C-D are adjacent frames of babble. Each frame is associated with a mixture in the aggregate speech model which are modeled by GMMs. Each mixture represents an acoustic region. For speech from a single subject as shown on the left, adjacent acoustic features (e.g., A-B-C, D-E-F, G-H-I) will have movement in the same acoustic region (e.g., A-B-C to mixture 1). For multispeaker babble, adjacent frames reside will move randomly across acoustic regions (e.g., A to mixture 3, B to mixture 2, and C to mixture 2). Therefore, it is expected that a measure of speaker babble can be obtained by determining how long we stay within a pdf over the time using a general GMM. If we hop frequently within mixtures for adjacent frames, there is greater spectral variation and we expect it to be babble. If consecutive frames appear to stay with the same GMM mixture longer, less spectral variability is present and it is more likely a single speaker. UBM is employed for analyzing the movement in the acoustic space. UBMs have been used for modeling background speakers for the speaker verification task [19]. A UBM is a GMM trained with speech streams from individual speakers. This represents an aggregate model for speech by a single speaker. If features from a single speaker are assigned to the largest scoring Gaussians in the UBM based on the maximum-likelihood criterion, contiguous blocks would reside in the same Gaussian. As the number of speakers increases, movement between acoustic regions should result for adjacent frames across babble data streams. The UBMs in our case are trained with speech from individual speakers, similar to the models used for speaker identification systems.

A. *Analysis of Acoustic Movement*

To demonstrate the impact of the number of speakers on the acoustic variability of babble, a 256-mixture UBM is constructed using all the training data from the TIMIT corpus. From this data, 19-dimensional MFCCs are extracted using a 20-ms window with a 10-ms skip between adjacent frames. Individual Gaussians in the UBM can be viewed as models of acoustically similar phone blocks in the training features. If the test audio stream contains speech from a single speaker, contiguous frames are expected to be acoustically similar, resulting in contiguous frames associated with the same Gaussian. As the speaker count in the babble increases, there is an increased hopping between Gaussians due to the acoustic variation in the data. To quantify the degree of abruptness in babble, a measure of the number of hops per audio segment frame of data is proposed. A hop is defined as a movement between Gaussians

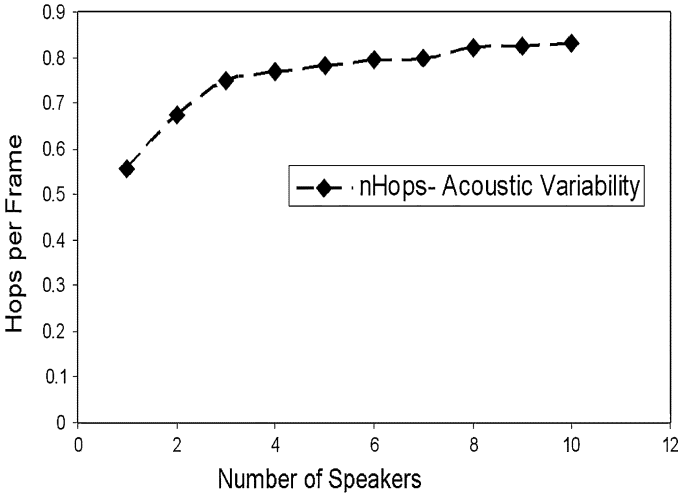


Fig. 11. As the number of speakers in babble increase, nhops increases due to decrease in frame contiguity.

in the UBM[2]. The average number of hops per frame is defined as

$$\text{meanhops} = \frac{\text{Total number of hops for the utterance}}{\text{Number of frames in the utterance}}. \quad (14)$$

The value of meanhops is between 0 and 1. If the value of nhops is 1, it implies that the average residence time for a frame in the Gaussian is 1 frame, which corresponds to every consecutive frame being associated with a different Gaussian. When nhops is 0.5, a single hop between Gaussians occurs every two frames. Fig. 11 shows the relation between the number of hops (meannhops) versus an increase in the number of speakers in the babble instance of duration 1 min. The average residency monotonically decreases (i.e., hops increase) with an increasing number of speakers in the babble. The relative change in meanhops is more for a smaller number of speakers (one to two speakers), as compared to when more subjects are in babble, where the babble is less time varying and nhops becomes constant. When the number of speakers approaches ∞, it is expected that the number of hops will reduce as babble will tend to be stationary. In the previous section, two aspects of babble were analyzed with the first being the shrinkage of the acoustic space as the number of speakers increase, and the second is the increased chaotic structure in babble with an increase in the number of participating subjects in babble. It is important to note the different time-domain analysis frame-lengths chosen for the two experiments.

These two aspects of babble are complementary; A decrease in acoustic volume indicates that with an increasing number of speakers in babble, the babble is less time varying in the long term, but for shorter segments the chaotic nature of babble increases. It should be noted that for analysis of the acoustic space, large frames of duration 125 ms are chosen versus 20 ms are chosen to assess durational continuity. Another observation from the second experiment is that UBMs constructed from speech utterances of individual speakers do not necessarily model the exact time varying nature of babble. Next, a system to detect the number of speakers is proposed based on the observation that the acoustic volume becomes concentrated as the number of speakers in babble increases. As observed in Section II, the number of speakers at any given time is approximately the number of conversations. Fig. 1 describes the construction of a two conversation babble audio stream. As shown in the figure, each stream consists of data from a single conversation. The babble stream from two conversations is constructed by overlapping individual conversations from Switchboard. In a babble data stream, the identity of the individual speakers is lost. Fig. 12 shows the histograms of frame count for a fixed number of speakers for two, four, six, and nine conversations. These histograms are oracle histograms constructed from the transcripts of the Switchboard corpus. Switchboard is a corpus of over 240 h of spontaneous telephone speech. It contains both A and B sides of telephone conversation, making it suitable to simulate babble conversations. Under the assumption that each conversation has only one speaker speaking at any point in time, the average number of speakers detected is equal to the number of conversations.

The pdf distributions for the number of speakers speaking per frame in babble is shown in Fig. 12. From the model for babble described in Section II, the number of conversations reflects the number instantaneous speakers in babble, under the assumption that there are two subjects participating in any single conversation, the total number of speakers would be twice the number of conversations. If the number of speakers speaking at a given instance is close to the number of conversations, detecting the number of speakers in babble requires a known relationship between the number of conversations in the acoustic environment and the number of speakers. The number of speakers speaking at a time is a function of the following variables:

- the topic of conversation;
- how the speakers provide input in the conversations (e.g., some speakers are active and contribute, while others are passive and spend more time listening).

Depending on the individual nature of each conversation, the resulting babble will take on many forms. As illustrated in Fig. 13, a two-stage detector for detecting the number of speakers at a given time is proposed. The first stage detector estimates the number of speakers for each frame. A speaker number histogram is generated for each frame in the data stream. This histogram is expected to have considerable fluctuation since the number of speakers active can vary from zero to the total number of participants in all conversations. In the second stage, the histogram is then considered as a feature, with its dimension N being a function of the maximum number of conversations to be detected (the maximum value of N is restricted by acoustic variability). Next, the histogram is normalized using the total number of frames in the data stream. This feature is seen to be highly correlated for a babble sequence. Finally, a discrete cosine transform DCT is applied and the first ten dimensions are

[2]Here, we assume that each Gaussian in the GMM corresponds to a unique phoneme. As the number of speakers in the UBM increases it is possible that more than one pdf will be used to represent the shoulder of a phoneme distribution

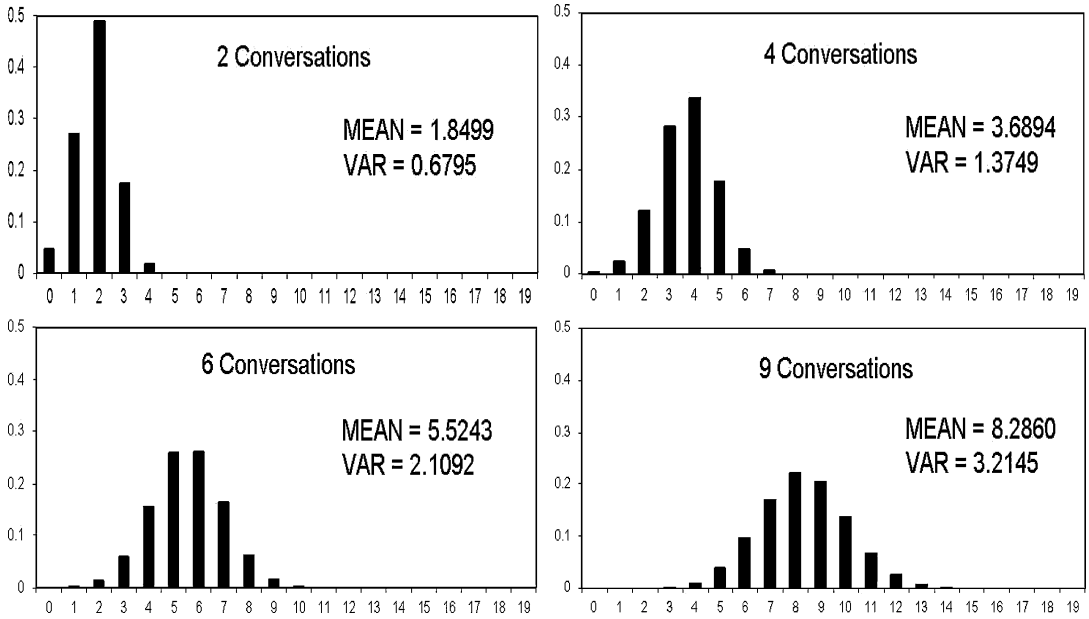


Fig. 12. PDFs for number of speakers per frame for babble constructed using two, four, six, and nine conversations, with an increase in the number of conversations the distribution has a larger variance.

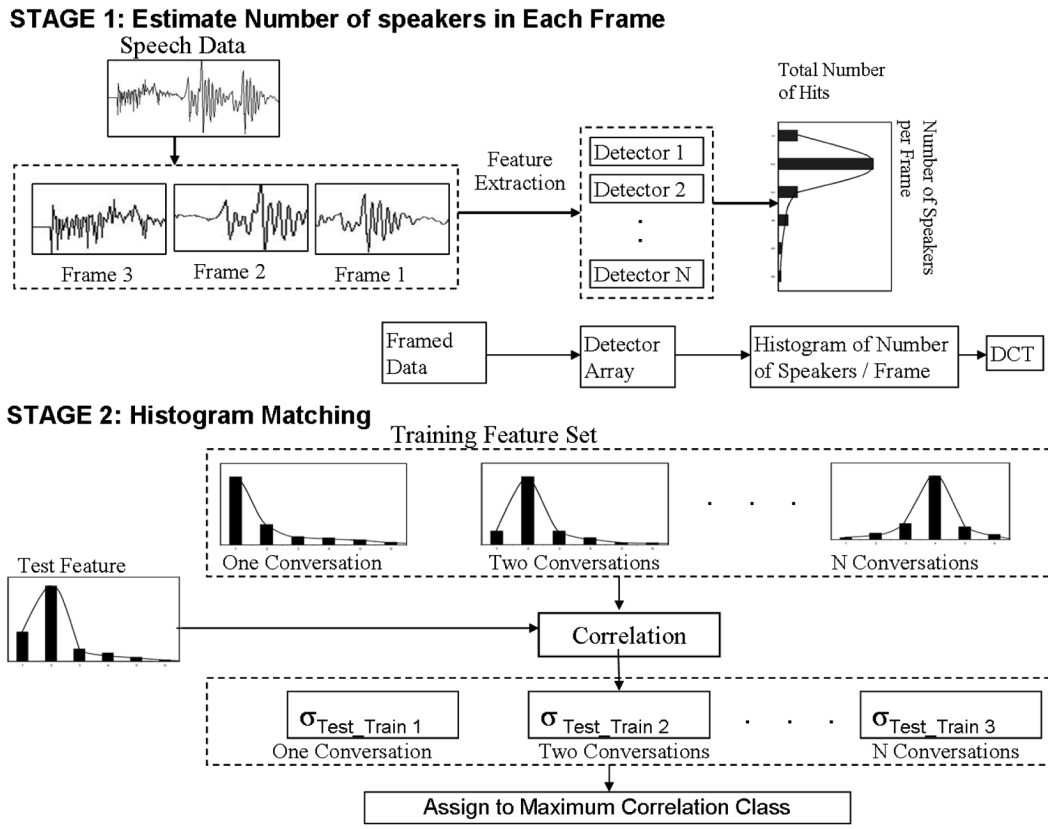


Fig. 13. Flow diagram for detecting number of speakers speaking at time.

employed for classification in order to reduce the dimensional correlation as well as reduce the data dimensionality.

V. Detection of Number of Speakers

A system is proposed for a closed set, where the maximum number of speakers speaking at a time is fixed to a number N. The detection scheme is a two stage detection scheme, where the preliminary detector decides on a preframe basis the number of speakers, and the second stage decides the number of conversations in an utterance. The second stage detector uses perframe decisions from the preliminary detector.

Let a set of training feature vectors be denoted by $X = \{x_1, x_2, \ldots, x_n\}$. Here, n denotes the number of frames in the training set. If Λ_k represents the model for babble with k speakers, then each frame x_i is classified according to the most likely number of speakers $\hat{\Lambda}$ as

$$\hat{\Lambda} = \arg \max_{1 \le i \le k} p(x_i \mid \Lambda_n). \tag{15}$$

Using the above decisions for all frames of an utterance, a probability mass function for the number of speakers in the given utterance is evaluated as follows:

$$P\{n = k\} = \frac{\{\text{total number of frames detected as having } k \text{ speakers}\}}{\text{total number of test frames}}. \tag{16}$$

A DCT of the observed pdf is evaluated. The DCT reduces the dimensionality of the feature vector and makes the dimensions independent. The DCT of this feature vector for k conversations is denoted by $G_k(z)$. Here, z is the dimension of the feature vector. The test feature $\hat{G}(z)$ is classified according to the following criterion:

$$k = \arg \max_{1 \le i \le k} \frac{G_k(z)\hat{G}(z)}{(\sigma_{G_k}\sigma_{\hat{G}})}. \tag{17}$$

Here, σ_G is the covariance of G. The test feature is assigned on the basis of the highest correlation. To implement the detection scheme, separate detectors for 1-to-N babble speakers are trained, and each test frame is assigned to one detector for every utterance. A hard speaker count decision is made on a per-frame basis. The first stage detector is trained using TIMIT data, since this data is read speech with limited pause sections within an utterance. This leads to a speaker count specific model for N speaker babble since read speech contains limited pause sections. The second stage uses a correlation-based detector. This proposed second-stage is required because in actual conversations, the number of speakers speaking at any given time can vary depending on the nature of the conversation. To train for a fixed number of speakers, babble samples with the required number of speakers are used as enrollment features. The training features are obtained from this enrollment feature data and averaged over the enrollment sequence to provide the train enrollment feature. After the test data feature extraction, the correlation of the test feature is measured across the closed set of enrollment features. The overall decision for the number of speakers for a given utterance is decided based on the maximum correlation with the test data.

VI. Results—Detection of Number of Speakers

As previously described, the speaker babble count detector consists of two stages, where each stage is presented separately below.

A. Stage 1: Preliminary Detector (Short Term)

The preliminary Stage 1 detector is made from babble data with an analysis frame length of 125 ms with no overlap between consecutive frames. For parameterization, 19-dimensional MFCCs are extracted as features. The resulting histograms of the babble speaker count detected from overlapped Switchboard corpus conversations is shown in Fig. 14. If we compare Fig. 12 with Fig. 14, it is observed that the detection performance is very poor for the correct number of speakers for a given frame. The detector output is skewed whereas the oracle pdfs are symmetric. It also is observed that the histograms vary with a change in the number of babble speakers. This feature is used to design the second stage detector.

B. Stage 2: Number of Speakers Detector (Long Term)

This stage of the framework is evaluated on simulated data using the Switchboard corpus for a total of 110 h of data constructed by overlapping different numbers of babble speakers to form each test/training utterance. The test and train were separate instances of babble with no overlap of the same speakers (i.e., the actual speakers used to create the overlapped babble speech were different for test and train). The data was split into a total of 800 utterances across nine test cases (each test case having N (from one to nine) conversations). The training set consists of 60 instances of babble for nine test cases. Babble data was framed using window lengths of 0.125 ms (1000 samples at 8 KHz). Results for babble speaker count classification of the number of speakers is shown in Table II. As the number of conversations increase, the acoustic separation in the data decreases and hence the error in detecting the exact speaker count increases (i.e., it is easier to detect the difference between three-to-five babble speakers versus 13 to 15 babble speakers, because the spectral diversity will decrease as the speaker count increases). On the other hand, the accuracy is very high for a speaker count between window ± 1 of the expected speaker count. This is expected in the probability distribution (Fig. 12) of the number of speakers when conversations overlap. From Table II, it is seen that the lowest babble speaker count performance with a detection window of ± 1 is about 81.6% when seven conversations are present. Given the nature of the task, it is difficult to accurately determine the actual number of people in a conversation at a given point of time, but by estimating the bounds on the number of conversations, it is possible to estimate the minimum number of people in babble.

VII. Babble Noise and Robust Speech Systems

To study the impact of characterizing the number of speakers in babble when babble is a primary source of additive noise, an in-set/out-of-set speaker verification system is employed (a full description of in-set recognition is found in [16]). For in-set speaker recognition, the test utterance is scored against all in-set speaker models relative to the background model. If the speaker is detected as any of the in-set models, the speaker is said to be an in-set speaker.

The primary motivation for this phase of the study is to determine the impact of choosing the correct babble speaker count in background for attempting to match the test and train background scenarios, and to study the impact of the ± 1 error in babble speaker count detection. To achieve this, the train and test speaker utterances are degraded with babble containing a different number of speakers. For a given signal-to-noise ratio (SNR), the closest corresponding matching (having a similar

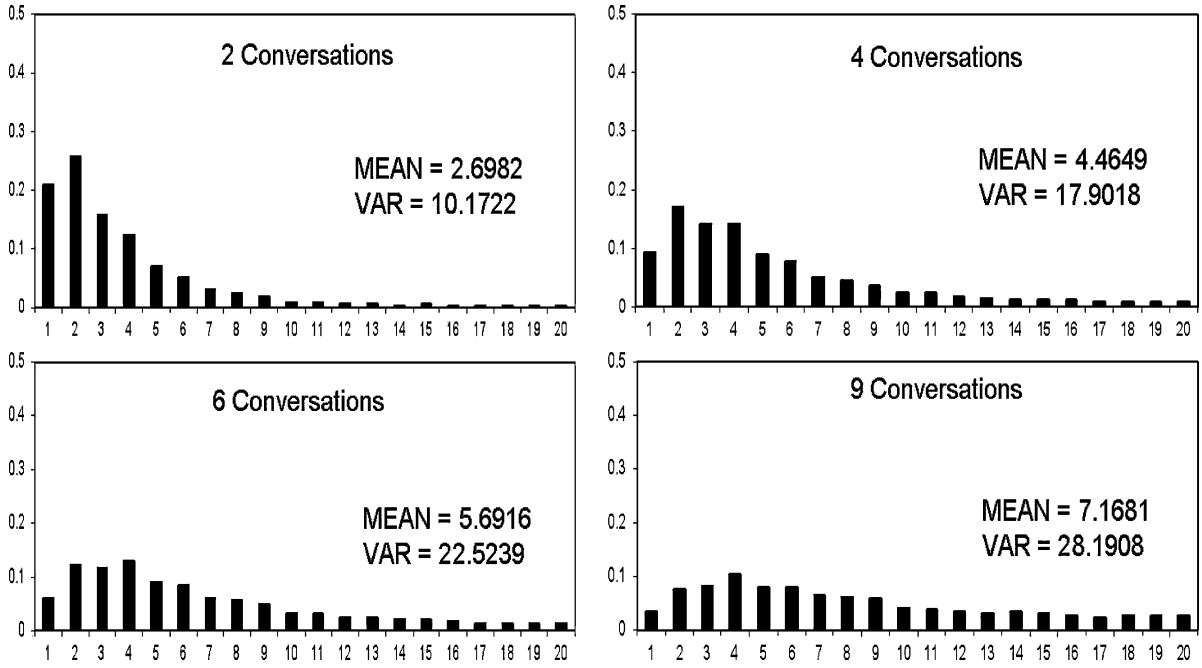


Fig. 14. PDFs for number of speakers per frame in babble when babble is constructed using two, four, six, and nine conversations.

TABLE II
CONFUSION MATRIX IN % OF THE NUMBER OF CONVERSATIONS DETECTED TO THE ACTUAL NUMBER OF CONVERSATIONS, EACH ROW CONTAINS THE CLASSIFICATION PERCENTAGES. THE LAST COLUMN CONTAINS DETECTION ACCURACY WITH A SPEAKER WINDOW SIZE OF ± 1

		DETECTED SPEAKER COUNT MODEL									AVG ACCU
TEST UTTERANCE BABBLE SPEAKER COUNT		1	2	3	4	5	6	7	8	9	
	1	**97.91**	**1.33**	0.76	0	0	0	0	0	0	**99.24**
	2	**31.25**	**65.33**	**9.87**	3.75	0	0	0	0	0	**96.25**
	3	0	**13.33**	**58.02**	**26.25**	3.65	0	0	0	0	**93.35**
	4	0	0	**18.51**	**56.25**	**19.51**	5.33	0	0	0	**94.67**
	5	0	0	1.23	**23.75**	**52.43**	**17.33**	6.66	1.35	0	**90.75**
	6	0	0	0	5	**28.04**	**44**	**14.66**	2.7	1.38	**86.7**
	7	0	0	0	0	8.53	**32**	**25.33**	**24.32**	8.33	**81.65**
	8	0	0	0	1.25	2.43	12	**22.66**	**18.91**	**41.66**	**83.23**
	9	0	0	0	0	0	2.66	9.33	**13.51**	**73.61**	**87.12**

speaker count babble) test models are chosen. Here, the speaker characterization is achieved on the basis of the number of speakers in babble as shown in Fig. 15. From the input data, the number of speakers in the background babble noise is estimated while keeping the SNR fixed, and the target models having the same number of speakers is chosen. The speaker verification system employs a binary detector that assigns a test token to the most likely in-set or out-of-set (UBM) model. The efficiency of this binary detector is measured in terms of equal error rate (EER). Here, the EER represents the classification error when the probability of false accept is equal to the probability of false reject. A lower EER indicates a better overall detection system, assuming equal cost for false reject and false accept. In general, when noise is introduced under matched test/train conditions, the EER increases. The next section describes experiments where the number of speakers in the babble noise is used to determine the in-set/out-of-set models to be used. Here, the attempt is not to improve performance for the in-set system, but to demonstrate that the selection of an adequately matched condition (in terms of the number of corrupting babble speakers) helps maintain overall performance. The next section describes the experimental setup.

VIII. EXPERIMENTS

A corpus of babble data is generated by varying the number of speakers in the babble. For a fixed number of speakers in babble, a corpus of ten babble instances is divided into sections of 3, 3, and 4 instances for test, train, and development respectively. Each of the babble instances are constructed using a different set of speakers (i.e., the exact speakers used for training, development, and testing are mutually exclusive). Each of the test, train, and development sets are degraded with their respective babble instances at a fixed SNR. The speaker ID system is evaluated over three conditions: for 15, 30, and 45 in-set speakers and for different duration of test utterances: 2, 4, 6, and 8 s, respectively. For a fixed SNR, a total of 12 conditions are evaluated. The in-set/out-of-set database consists of the male speakers for the TIMIT corpus at 16 kHz, and babble distortion is constructed using female speakers from TIMIT. The features used for classification are 19-dimensional MFCCs. Babble is modeled as a

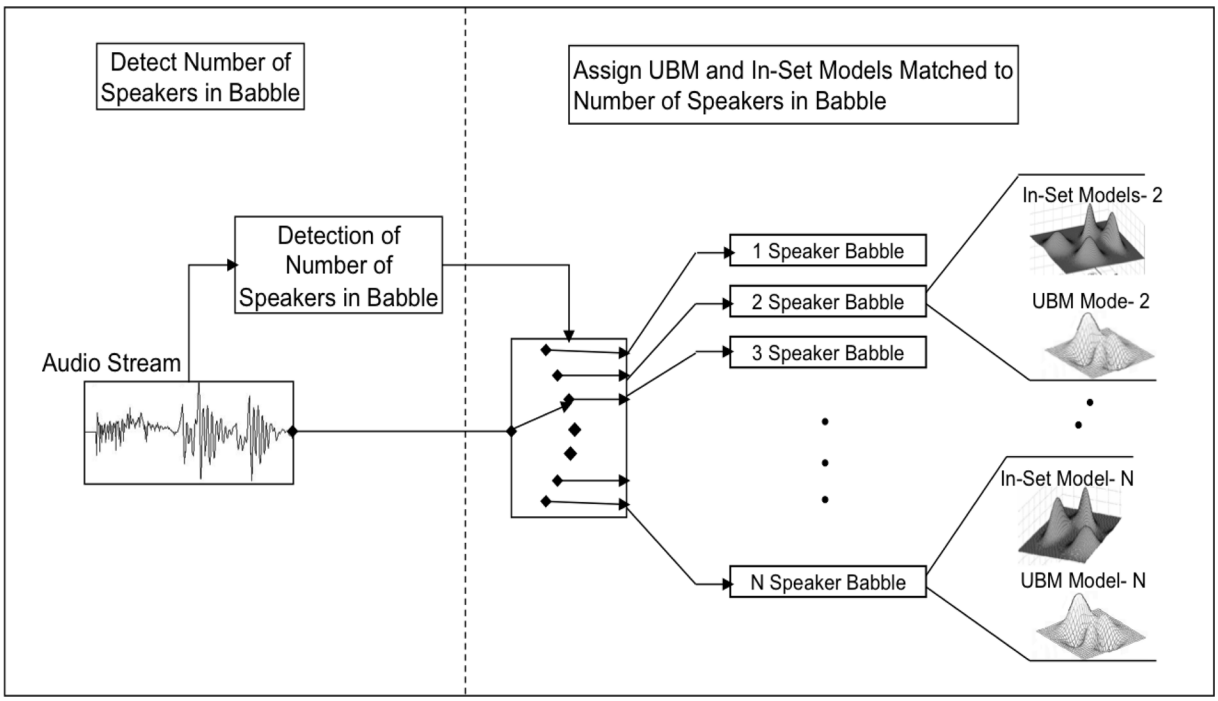


Fig. 15. Schematic for using the number of speakers for maintaining performance for in-set/out-of-set speaker verification.

TABLE III
BASELINE IN-SET/OUT-OF-SET SPEAKER RECOGNITION SYSTEM PERFORMANCE IN CLEAN CONDITIONS

Test Duration	2sec	4sec	6sec	8sec
15in-set EER	11.019	7.222	6.389	5.556
30in-set EER	15.556	10.278	6.667	6.667
45in-set EER	14.352	9.815	8.056	9.444

function of the number of speakers speaking at a time. The next section evaluates the performance of the speaker verification system where detected babble noise information is incorporated within the system.

IX. RESULTS

The performance mismatch was evaluated for babble noise where the speaker count varied from 1–10 speakers. The UBM was trained using 60 male speakers from the TIMIT corpus which are separate from the in-set/out-of-set speakers. Here, results are presented for speech degraded at 10 dB, though results are similar for different SNRs (e.g., 5–15 dB). Training data for each in-set speaker is about 5 s. Table III shows the baseline performance of the in-set speaker verification system without the introduction of babble distortion. The average performance of the speaker verification system under different clean conditions is 9.25% EER. For speech corrupted by babble noise, where the number of speakers in babble varying from one to nine at 10-dB SNR under matched test/train conditions, the performance drops to 27.94% EER. Test conditions are considered to be matched when the speaker count in babble is within a babble speaker count window of ± 1 of the actual speaker count. Mismatch is present when models are chosen outside of this ± 1 babble speaker window. Performance mismatch for each speaker number condition is evaluated using the relation

$$\text{mismatch} = \sum (\text{EER}_{\text{exact}} - \text{EER}_{\text{model}})/N. \qquad (18)$$

This mismatch is the average performance mismatch between the exact EER and the EER when a different model is chosen as the target model. Table IV shows the average performance mismatch under matched and mismatched conditions for the task. As observed, matched cases always outperform the mismatched condition. Also, performance with a reduced number of subjects in the degrading babble is better than performance when a model with more number of speakers in babble is used. The average performance mismatch across all conditions when matched models are chosen is shown in Table V. The EER performance loss under matched conditions (± 1 speaker difference in selected babble noise model) is -0.42% as compared to an average -6.44% EER loss when models are chosen outside this ± 1 window size. This corresponds to an average 23% relative improvement on the EERs across all conditions by choosing the appropriate set of in-set/out-of-set speakers plus babble noise models. Therefore, employing babble noise model detection helps maintaining overall speaker ID performance.

Another observation is that it is better to choose speaker models with a reduced number of speakers in the babble. This can be attributed to the difference in background speakers aiding the separation in the speaker ID system. With an increase in number of speakers, the test and training instances of babble are not as distinguishable and this reduces the background contributing to the speaker separation. The babble model detector influence is more important as the number of speakers in babble increases.

TABLE IV
PERFORMANCE MISMATCH OF THE IN-SET SPEAKER RECOGNITION SYSTEM UNDER MATCHED AND MISMATCHED BABBLE NOISE CONDITIONS FOR EACH TEST DURATION. THE EXACT EER IS WHEN TEST AND TRAIN UTTERANCES ARE DEGRADED WITH BABBLE HAVING THE SAME NUMBER OF SPEAKERS. THE NEXT ROW (MATCHED CONDITION ± 1) SHOWS THE AVERAGE EER PERFORMANCE DIFFERENCE WHEN MODELS HAVING ± 1 NUMBER OF SPEAKERS IN BABBLE NOISE ARE CHOSEN THE LAST ROW (MISMATCHED CONDITION) FOR EACH TEST DURATION SHOWS THE PERFORMANCE DIFFERENCE WHEN MODELS OTHER THAN THOSE HAVING "SPEAKER COUNT" IN THE VICINITY OF ± 1 ARE CHOSEN

	NUMBER OF BABBLE SPEAKERS IN NOISE							
	2	3	4	5	6	7	8	9
2sec Exact EER	**29.91**	**28.75**	**30.60**	**30.97**	**31.57**	**32.64**	**31.67**	**32.41**
Δ EER: Matched Condition	*2.08*	*-1.97*	*-0.74*	*-0.28*	*-0.09*	*+1.53*	*-1.60*	*+0.19*
Δ EER: Mismatched Condition	*-3.17*	*-7.87*	*-5.74*	*-6.59*	*-6.16*	*-6.33*	*-9.12*	*-11.94*
4sec Exact EER	**22.87**	**23.98**	**24.26**	**24.54**	**23.33**	**27.78**	**25.46**	**23.98**
Δ EER: Matched Condition	*-0.56*	*-0.56*	*+0.19*	*+0.56*	*-2.64*	*+1.07*	*-1.16*	*-0.83*
Δ EER: Mismatched Condition	*-4.31*	*-5.33*	*-4.78*	*-6.63*	*-6.17*	*-3.76*	*-6.63*	*-11.79*
6sec Exact Condition	**17.50**	**21.94**	**23.89**	**21.94**	**20.28**	**24.58**	**25.14**	**20.56**
Δ EER: Matched Condition	*-2.08*	*+0.35*	*+0.76*	*-1.04*	*-4.17*	*+0.35*	*+1.94*	*-3.33*
Δ EER: Mismatched Condition	*-4.91*	*-2.53*	*-2.03*	*-5.08*	*-6.78*	*-5.36*	*-4.78*	*-11.23*
8sec Exact Condition	**18.61**	**17.50**	**21.67**	**22.50**	**20.00**	**20.56**	**23.61**	**18.06**
Δ EER: Matched Condition	*+3.33*	*-5.42*	*+0.83*	*+1.39*	*-1.39*	*-1.53*	*+2.22*	*-6.11*
Δ EER: Mismatch Condition	*-0.46*	*-3.83*	*-1.67*	*-2.22*	*-5.11*	*-8.78*	*-5.06*	*-12.73*

TABLE V
PERFORMANCE OF THE INSET SPEAKER RECOGNITION SYSTEM UNDER MATCHED AND MISMATCHED BABBLE NOISE CONDITIONS FOR EACH TEST DURATION. THE SNR FOR BABBLE NOISE IS 10-dB BL (BASELINE WITH 10-dB-BABBLE NOISE INTRODUCED). BASELINE EER% IS SHOWN FOLLOWED BY ΔEER FOR ME

In-Set / Out-of-Set	TEST / TRAINING UTTERANCE DURATION											
	2sec			4sec			6sec			8sec		
	BL	ME	MME	BL	ME	MME	BL	ME	MME	BL	ME	MME
15 / 45	**31.22**	*-0.1*	*-7.11*	**24.79**	*-0.49*	*-6.17*	**22.22**	*-0.9*	*-5.33*	**20.92**	*-0.83*	*-4.98*
30 / 30	**35.72**	*+0.17*	*-7.78*	**31.11**	*+0.26*	*-6.6*	**28.61**	*-0.18*	*-6.61*	**25.83**	*-1.09*	*-7.79*
45 / 15	**34.3**	*-0.06*	*-7.38*	**29.01**	*-0.24*	*-6.59*	**27.37**	*-0.12*	*-4.37*	**24.25**	*-1.49*	*-4.36*

X. FUTURE WORK AND IMPROVEMENTS

This study has considered the problem of analysis, modeling, and detection of characteristics of babble speech, known to be the most challenging noise interference in speech systems today. There are significant differences between babble collected from real speaker scenarios and babble constructed by adding individual speaker streams of data together. The differences arise due to different data acquisition channels, when data is collected from individual speakers or there are conversations collected from close microphones. In contrast, when babble is collected in natural settings (example in a meeting room scenario) a far-field microphone is used. This leads to significant differences in channel conditions. The impact of the language of babble in different speech systems, and the ability to detect the particular languages of the babble is currently under study. Finally, the impact of group stress/emotion on babble and its impact on speech systems is an interesting field for further investigation.

XI. CONCLUSION

In this paper, a framework to characterize babble noise is proposed. Babble is known to be the most challenging distortion in speech systems, due to its speaker/speech like characteristics. There are differences in the number of speakers per frame pdfs when babble noise is modeled as a sum of conversations as opposed to adding individual streams of speakers. One of the main factors impacting the nature of babble is the number of speakers in babble noise. An algorithm was proposed to detect the number of speakers in a given instance of babble. The algorithm was evaluated on simulated conversations from the Switchboard corpus. Detection performance of over 80% accuracy is obtained in detecting speaker count to within ± 1 of the number of conversations, given that each conversation is assumed to be consisting of two speakers. The performance is encouraging, given the significant challenge in characterizing babble speech. It is believed that this represents one of the first studies to specifically address the underlying structure of babble noise. This finding from characterization of babble opens up possibilities for future work and also impacts existing applications. Babble can be used as a source of information (language ID, gender ratio, group emotion characteristics, etc.) itself. In our data collection, we have found different babble characteristics when the previous parameters have changed. This information can be of value in and of itself for the purposes of environment forensics. Alternatively, this information can be used in order to supplement speech systems in order to maintain performance in the most challenging of noise types. Here, the impact of babble noise on speaker verification has been studied, where the impact of babble speaker count detection was shown to help overall performance. It has been shown that proper selection of in-set speaker plus babble noise models can improve

the performance of in-set/out-of-set speaker verification by 24% compared to choosing a generic babble model. One drawback of the current setup is that it requires a sufficient data for characterizing the number of speakers. Second, the work has been primarily focussed on modeling babble as the number of speakers; such a modeling suffices for speaker identification systems, but for speech recognition additional information such as language information of the background is required. This is important because English speech recognition in English babble would be more challenging than English speech recognition in background babble consisting of a foreign language. It is suggested that these initial findings will open a scope of innovation and applications in the study of babble for speech and language technology.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, pp. 261–291, Apr. 1995.

[2] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[3] J. H. L. Hansen and D. Cairns, "ICARUS: A source generator based realtime system for speech recognition in noise, stress, and Lombard effect," *Speech Commun.*, vol. 16, no. 4, pp. 391–422, Jul. 1995.

[4] J. H. L. Hansen and V. Varadarajan, "Analysis and normalization of Lombard speech under different types and levels of noise with application to in-set speaker id systems," *IEEE Trans. Audio, Speech, Lang. Process.*, to be published.

[5] M. Akbacak and J. H. L. Hansen, "Environmental Sniffing: Noise knowledge estimation for robust speech systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 465–477, Feb. 2007.

[6] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, 1990, pp. 845–848.

[7] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 3, no. 119, pp. 1562–1573, Mar. 2006.

[8] G. Li and M. E. Lutman, "Sparseness and speech perception in noise," in *Proc. Interspeech*, Pittsburg, PA, 2006, pp. 1466–1469.

[9] P. C. Loizou*, Speech Enhancement: Theory and Practice.*. Boca Raton, FL: CRC, 2007.

[10] N. Morales, L. Gu, and Y. Gao, "Adding noise to improve noise robustness in speech recognition," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 930–933.

[11] P. D. O'Grady, A. B. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *Int. J. Imag. Syst. Technol.*, vol. 15, pp. 18–33, Aug. 2005.

[12] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, no. 1, pp. 151–173, Nov. 1996.

[13] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 28, no. 4, pp. 367–376, Aug. 1980.

[14] J. R. Deller, J. H. L. Hansen, and P. J. Proakis*, Discrete-Time Processing of Speech Signals.*. New York: Wiley, 1999.

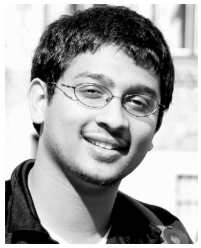
[15] L. R. Rabiner and R. W. Schafer*, Digital Processing of Speech Signals.*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[16] V. Prakash and J. H. L. Hansen, "In-set/out-of-set speaker recognition under sparse enrollment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2044–2052, Sep. 2007.

[17] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," in *Proc. Interspeech*, Philadelphia, PA, 1996, vol. 1, pp. 466–469.

[18] Y. Gong, "Stochastic trajectory modeling and sentence searching for continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 33–34, Jan. 1997.

[19] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.



Nitish Krishnamurthy received the B.E. degree in instrumentation and control engineering from the University of Pune, Pune, India, in 2004 and the M.S. degree in electrical engineering from the University of Texas at Dallas, Richardson, in 2007. He is currently pursuing the Ph.D. degree at the Center for Robust Speech Systems, University of Texas at Dallas.

He has been a Research Intern at Texas Instruments in the area of speech and language systems, during the summers of 2007 and 2008. His research focuses on acoustic noise characterization for speech systems. His research interests also include embedded speech to speech translation and speech recognition systems.

John H. L. Hansen (S'81–M'82–SM'93-F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1988 and 1983, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is a Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as a Professor in the School of Brain and Behavioral Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 43 (18 Ph.D., 25 M.S.) thesis candidates, is author/coauthor of 294 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior Modeling* (Springer, 2008), and lead author of the report "The impact of speech under 'stress' on military speech technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was the recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body. He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and is serving as Technical Program Chair for IEEE ICASSP-2010, Dallas, TX. In 2007, he was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise," and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee. Previously, he served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005–2006), Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–2003). He has also served as a Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the ISCA (International Speech Communications Association) Advisory Council.