Babble Speech: Acoustic and Perceptual Variability

Nitish Krishnamurthy, Ayako Ikeno, and John H. L. Hansen

Center for Robust Speech Systems (CRSS),
Department of Electrical Engineering, Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, TX, U.S.A.
nkm052000@utdallas.edu, ikeno@utdallas.edu, john.hansen@utdallas.edu

Abstract

The presence of babble noise is one of the most difficult environments to sustain speech system performance. This study focuses on acoustic and perceptual analyses of babble. The acoustic variability of babble is analyzed as a function of the number of speakers in babble. The concept of acoustic volume is proposed and it is shown that the acoustic volume reduces as the number of speakers in babble increase. This framework is evaluated using over 40 hours of simulated babble from SWITCHBOARD corpus. It is observed that the acoustic volume does not change significantly when there are 4 or more speakers in babble. It is also seen that with an increase in the number of speakers in babble, there is an uneven spread of data in the acoustic space due to mixing of multi-speaker content. This study provides a framework to better understand the babble environment, enabling us to improve the formulation of reliable algorithms for robust speech systems.

Index Terms: Babble, Robustness, Speech System

1. Introduction

One of the most frequently encountered and challenging environments for speech systems is multi-speaker babble. Babble noise is the interference from background speech sounds in the acoustic vicinity of target speech. The similarity to the target speech and the non-stationary nature makes it an especially difficult noise environment for all speech systems. Any mobile speech solution would be incomplete if it does not effectively address the issue of babble speech which can result in serious loss of performance. Moreover, the study of babble can provide additional information about the environment which can be used for various applications.

Previous research [1], [2], [3] examined perception of competing/overlapping speech sounds. Areheart [1] compared perception of competing vowels for normal and hearing impaired subjects. Assmann [2] proposed time space models to better model perception of competing vowels. Loizou [3] reported that an increase in the number of speakers in babble results in reduced ability to recognize monosyllables from individual speakers. As the number of speakers increases, the ability to recognize monosyllables from individual speakers decrease. Krishnamurthy and Hansen [4] made the first attempt to model babble and formulated a framework to detect the number of speakers in babble. Babble is affected by the acoustic environment, language, gender, and the number of participating speakers etc.. The most important factor dictating the nature of babble is the number of speakers, and hence conversations, since it decides the contribution for the remaining characteristics.

This paper focuses on the study of variability of babble speech and its dependence on the number of speakers has been considered. It is shown that an increase in the count of speakers participating in babble, results in babble with less time variation, reducing our ability to extract information effectively. This study aids the understanding of acoustic properties of babble speech. It is indicated that the rate of change in acoustic characteristics of babble is very slow when the number of speakers in babble is four or greater in the experiments. There is a greater difference between babble from 2 speakers vs. 3 speakers whereas, there is little change in babble when the speaker count increases from 5 to 6. The next section focuses on an analysis of babble as a series of overlapping phones.

2. Babble as a Sequence of Overlapping Phones

Babble can be viewed as a sequence of overlapping phones. Individual phones from each speaker overlap in the acoustic space giving unique properties to babble. As the number of overlapping phones increase within a given babble speech, the differences between individual phones are averaged out and their identity becomes blurred. This removes the ability to distinguish individual phones in a babble utterance. This impact of overlapping the phones on the resulting sounds is demonstrated using the reduction in Itakura-Saito(IS) distance [5]. This IS distance reduces when distinct phones are superimposed. The symmetric IS measure is shown in the following equation (1).

\[ d_{SI(j)} = \frac{1}{2}(d_j(\hat{a}_s, \hat{a}_s) + d_j(\hat{a}_s, \hat{a}_s)) \] (1)

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

\[ d_j(\hat{a}_s, \hat{a}_s) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |v(\omega) - v(\omega)| \, d\omega \] (2)

where,

\[ v(\omega) = \log \left( \frac{\sigma_{v_j}}{|A(\hat{a}_s, \omega)|^r} \right) - \log \left( \frac{\sigma_{v_j}}{|A(\hat{a}_s, \omega)|^r} \right) \] (3)

The experiments are conducted using synthetic phones from the same speakers generated by the Festival Speech Synthesizer system. The phones generated are /@l/, /A/, /Y/, /U/, /i/ at 16 Khz for 12ms. Table 1 presents samples of IS distance when
different phones are overlapped. These waveforms are modeled using 12th order linear prediction coefficients (LPCs) [6]. Fig. 1 shows the mean IS distances and the variance of those distances as a function of number of overlapping phones. For a given count, various combinations of 5 phones are chosen and superimposed. There is a monotonic decrease in the mean and variance of the distances between the averaged phones as the number of phones in babble speech increases from 1 to 5. This suggests that, with an increase in the number of speakers in babble, the noise becomes localized in the acoustic space. A similar phenomenon is observed when listeners are subject to overlapped phone sounds. This is described in the next section.

2.1. Perceptual Clustering of Overlapped Phones

Speech samples produced by an American native English speaker were recorded in an ASHA (American Speech-Language-Hearing Association) certified sound booth for this study. The speaker was 23 year old, male graduate student from Minnesota currently at University of Texas, Dallas (UTD). Utterances were composed of single words - "the vowel. The listeners were asked to listen to each test sample and determine to which part of the chart (Fig. 2) the sound belongs.

2.2. Perceptual Experiment Results

As illustrated in Fig. 3, the results show that, when two or more vowels overlap, the probability of perceived vowels is higher in certain area of the vowel space. That is, toward the center (/x/ in the word cut) through low front vowel (/i/ in the word cut). This trend indicates that perceptual distance among vowels reduces with overlapping speech samples, and therefore supports the observation reported in the previous section. In the next section the concept of the clustering babble in acoustic space is formalized using the concept of Acoustic Volume.

3. Babble and Acoustic Volume

In the previous section, it was observed that as the number of overlapped vowels increase the resulting IS distances decreases. This observation can also be extended to general acoustic spaces. Let $X_1, \ldots, X_k$ 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 cuboid is used to model the acoustic space enclosed by these centroids. Fig.
Figure 4: Illustration of the acoustic area/volume occupied by a GMM of 4 Mixtures.

4 describes the construction of this space in two dimensions. The vertices of the above figure are given by \((x_{\text{max}}, y_{\text{max}}), (x_{\text{min}}, y_{\text{max}}), (x_{\text{min}}, y_{\text{min}}), \) and \((x_{\text{max}}, y_{\text{min}})\). 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,

\[
\{ \arg \max x_1, \arg \min x_1, \ldots, \arg \max x_N, \arg \min x_N \}.
\]

Here, the maxima and minima are evaluated for each dimension separately across all centroids. The entire acoustic space of the data is enclosed within these extremities. 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. This volume is therefore an indicator of the acoustic variation of the data. The volume of this enclosed \(N\) dimensional cuboid with adjacent edges \(e_1, e_2, e_3, \ldots, e_N\) is given by,

\[
V = e_1 * e_2 * e_3 * \ldots * e_N
\]

where,

\[
e_1 = \arg \max x_1 - \arg \min x_1.
\]

These distances are calculated for all centroids describing the acoustic space. For \(N\) centroids, there would be \(\binom{N}{3}\) distances. The analysis of the inter-centroidal distances therefore provides additional information on the distribution of the data (i.e., information pertaining to relative closeness of the centroids in the acoustic space).

4. Evaluation Setup

The volume and acoustic space characteristics are evaluated on a synthetic babble corpus constructed using the test corpus of TIMIT and a total of 40 hours of babble speech constructed using SWITCHBOARD(SWB). Babble constructed using TIMIT and SWB are different since monologues are overlapped in the first case whereas, in the second case conversations are overlapped. For the 2 speaker babble case, TIMIT babble would contain speech from 2 subjects. However, SWB babble would contain speech from 4 subjects with mostly 2 subjects speaking simultaneously. These speakers will change over time because of turn-taking. The babble from SWB is more realistic as shown in [4]. The analysis however, holds for both cases. The number of speakers are varied from a single speaker up to 10 speakers/conversations simultaneously. Here, 19 dimensional Mel Frequency cepstral coefficients(MFCC’s) are extracted from 125ms frames. The frame size has been chosen to be large to facilitate the analysis of the long term structure of babble. These MFCC vectors are assumed to characterize the acoustic space of babble by clustering and employing Gaussian Mixture Models (GMM). The GMM’s are given by the equation,

\[
p(x) = \sum_{j=1}^{N} w_j N(x | \mu, \sigma)
\]

where, \(N\) is the 19 dimensional Gaussian. The GMM model parameters are estimated using the EM algorithm, where the data is split into 64 mixtures and the means of each mixture is used to characterize the acoustic space. The acoustic volume is evaluated using these centroids.
5. Results

Fig. 6 shows a resulting monotonic decrease in the acoustic volume as the number of speakers in babble increases. As observed from the inset plot, there is little variation in the acoustic volume when there are more than 4 subjects in babble. Also noticeable is the fact that for a single speaker, data from SWB has consistently greater acoustic volume than for TIMIT. This observation can be used to conclude that the volume/variability gets richer when there is turn-taking in the conversation as compared to combination of monologues. The mean volume reduces almost exponentially with an increase in the number of speakers speaking simultaneously. This leads to the conclusion that babble data clusters in the acoustic space. In general, to process speech in noise, noise should be localized in this space and separated from the acoustic region of the target speech. However, most of the time, 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 subspace of the acoustic space occupied by speech from a single speaker. This is also shown by the perceptual experiments where subjects kept the synthesized speech within the region marked by the pivots.

Fig. 7 shows the pdfs of distances between the centroids for speech from 1 speaker, 4 speakers, and 8 speakers. As illustrated in figures below the pdf plots, with an increase in the number of speakers in babble, the distribution of data is skewed. This is expected, as adding speakers averages out the differences in the data, causing the data to cluster. The distance histograms with 1 speaker is most uniform, with movement towards a Gamma distributions as the number of speakers increases. With an increase in the number of speakers, the mean distance between the centroids decreases, implying the clustering of acoustic features. As evident from the volume and distance plots, cases where there are fewer number of speakers in babble, the centroids enclose a larger volume, and they become uniformly distributed. With an increase in the number of speakers, the average distance between the centroids reduce, the centroids begin clustering, and the enclosed volume decreases.

6. Conclusion

In this study, acoustic perceptual analyses of babble is performed. The acoustic variability of babble is examined as a function of the number of subjects participating in babble. The results showed that acoustic and perceptual vowel space reduces when phones overlap. That is, the overlapping sounds cluster around a particular region. This leads to skewed pdfs for intercentroidal distances. It was observed that when the number of speakers in babble is 4 or greater, the acoustic volume does not vary much. This study helps better understand the properties of babble and its localization in acoustic space, and this knowledge would help explain the change in performance for ASR engines in babble.

7. References


