



# Exploring subsegmental and suprasegmental features for a text-dependent speaker verification in distant speech signals

B. Avinash<sup>1</sup>, S. Guruprasad<sup>2</sup> and B. Yegnanarayana<sup>1</sup>

<sup>1</sup> International Institute of Information Technology, Hyderabad, India

<sup>2</sup> Department of Computer Science and Engineering, Indian Institute of Technology Madras, India

avinashb@research.iiit.ac.in, gurus@cse.iitm.ac.in, yegna@iiit.ac.in

## Abstract

Existing automatic speaker verification (ASV) systems perform with high accuracy when the speech signal is collected close to the mouth of the speaker (< 1 ft). However, the performance of these systems reduces significantly when speech signals are collected at a distance from the speaker (2-6 ft). The objective of this paper is to address some issues in the processing of speech signals collected at a distance from the speaker, for text-dependent ASV system. An acoustic feature derived from short segments of speech signals is proposed for the ASV task. The key idea is to exploit the high signal-to-noise nature of short segments of speech in the vicinity of impulse-like excitations. We show that the proposed feature yields better performance of speaker verification than the mel-frequency cepstral coefficients (MFCCs). In addition, regions of high signal-to-reverberation ratio, duration and pitch information are used to improve the performance of the ASV system for distant speech.

**Index Terms:** automatic speaker verification, text-dependent, distant speech, signal-to-noise ratio, pitch, duration.

## 1. Introduction

Speaker recognition is a generic term that refers to any task which discriminates people based upon their voice characteristics [1]. Although low error rates have been achieved for speaker recognition systems using close-speaking speech signals, the same is not the case for speech signals collected at a distance. Speech signals collected at a distance are affected by noise and reverberation, and the short-time spectral features derived from the segments of distant speech differ from the features derived from the corresponding segments of close-speaking speech. This leads to poor performance of speaker verification systems in the case of distant speech signals. However, human beings are able to identify speakers even at a distance. This suggests that some speaker-specific features are indeed preserved in distant speech signals.

The issues involved in performing speaker recognition using distant speech signals have been addressed in the literature using three broad approaches, namely, (a) compensation of spectral features, (b) enhancement of distant speech, and (c) use of multiple microphones for recording speech signals. In [2], a reverberation compensation method for a Gaussian mixture model (GMM) based speaker identification system was proposed along with a multiple channel combination and feature normalization scheme. In another study, a beam-forming enhancement was used to improve a single channel GMM-based identification system [3]. In [4], reverberant speech was synthesized from close-speaking speech, to train speaker models for reducing channel mismatch. Microphone arrays were also used

to improve the performance of a speaker recognition system [5].

All the above approaches have some limitations. The methods for enhancement of distant speech signal depend on the room transfer function of the recording ambience. This limits the portability of the speaker verification system, which is also the case with the use of multiple microphones. Compensation of features derived from distant speech is also dependent on the recording environment. Thus, there is a need for features which are robust to distance, and which retain the speaker-specific information at the same time. This paper presents an approach for text-dependent speaker verification using distant speech, by exploiting the robustness of subsegmental (1-3 ms) and suprasegmental (> 100 ms) features. In particular, spectral features are derived from short segments (2-3 ms) of speech signal, in the regions of high signal-to-noise ratio in each pitch period. In addition, suprasegmental features such as pitch and duration are exploited to enhance the performance of the system.

The paper is organized as follows: In Section 2, the database used for the study is described. Section 3 gives a description of the baseline system. In Section 4, extraction of a robust short-segment spectral feature is discussed. Section 5 explains the use of suprasegmental features to improve the performance of the system. Section 6 concludes with a summary of the key ideas proposed in this work.

## 2. Acquisition of distant speech

Speech signals were collected from 45 speakers in a laboratory environment with room dimensions of 20 ft × 15 ft × 10 ft. Speech was collected from 30 male and 15 female speakers. The microphones were placed at distances of 0.2 ft, 2 ft, 4 ft and 6 ft from the speaker. Four omni-directional microphones were used, which were placed such that the speaker is at the same horizontal level as the microphone setup. An illustration of the microphone setup is shown in Fig. 1. The points 1, 2, 3 and 4 correspond to the close (0.2 ft), 2 ft, 4 ft and 6 ft microphone positions respectively. The speaker was positioned just in front of the close-speaking microphone. Speech signals were collected simultaneously at all the distances, which were recorded at 48000 Hz and stored at 16 bit samples. Speech signals were resampled to 8000 Hz for processing. The sentence used as the text during recording is "We were away a year ago". The data were collected in three different sessions over a span of 7 months. Five repetitions of the sentence were recorded in the first session, while ten repetitions were recorded in both the second and the third session. Thus, a total number of 25 repetitions of the sentence was recorded by each speaker. Three utterances of session 1 were used for enrollment. Twenty utterances of session 2 and session 3 were used for verification. So, the to-

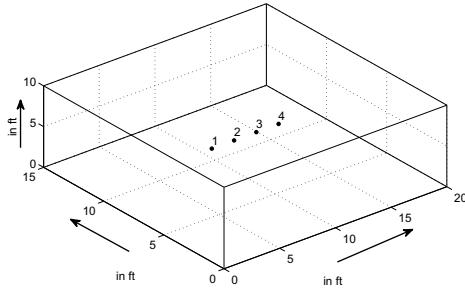


Figure 1: Microphone setup used in the data collection process.

tal number of genuine tests for 45 speakers is 900 ( $20 \times 45$ ). A speaker can be used as an imposter for the remaining 44 speakers. So, the number of imposter tests is 39600 ( $20 \times 45 \times 44$ ).

### 3. Baseline speaker verification system

A baseline text-dependent speaker verification system is developed using mel-frequency cepstral coefficients (MFCC) as features. The enrollment phase uses three utterances of a speaker to create three models for each speaker. The verification stage compares a test utterance with the three reference models of the claimant along with ten background models.

The speaker verification system consists of three stages: Feature extraction, pattern comparison and decision logic. A begin-end detection is performed for all the speech utterances before extracting the features. The baseline system uses 20 dimensional MFCC features extracted for each frame with a size of 20 ms and a shift of 5 ms. To reduce the effect of channel variations, cepstral mean subtraction (CMS) is performed after extracting the coefficients [6]. Dynamic time warping (DTW) is used for comparison of two sequences of MFCC features, which are of different lengths. Each test utterance is compared with three claimant models and ten background models. For the computation of matching score ( $\alpha$ ), only those frames along the optimal warping path are considered where the Euclidean distance between the test utterance and the claimant model is less than that between the test utterance and the background models.

The performance of the speaker verification system is evaluated as follows: All the genuine and imposter scores are normalized in the range of 0 to 1 such that a higher score indicates a greater possibility of acceptance.

The cost function  $\eta_C$  used for the system is given by

$$\eta_C = 0.1 \times \eta_R + 0.9 \times \eta_A, \quad (1)$$

where  $\eta_R$  and  $\eta_A$  denote the false rejection and the false acceptance rates, respectively. The threshold that yields the least cost for close-speaking speech signals is chosen as the threshold of the baseline speaker verification system. The same threshold is used for the speech signals collected at distances of 2 ft, 4 ft and 6 ft. The first row in Tables 1 and 2 shows the genuine acceptance and imposter rejection rates of the baseline system for different distances. Due to variation of the acoustic feature with distance, there is decrease in the confidence scores, which results in a reduction of the performance of the system. Since the same threshold is used for all the distances with the objective of reducing false acceptance, the speaker verification system progressively reaches a state where it rejects all the inputs. So,

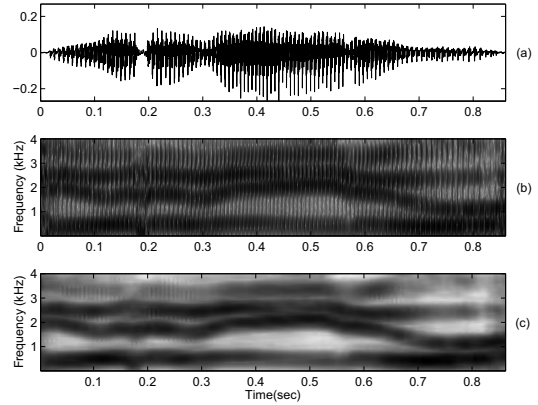


Figure 2: Illustrating the significance of short-segment analysis: (a) Close-speaking speech signal, (b) wideband spectrogram of the speech signal, and (c) wideband spectrogram obtained after time-averaging the short-time spectra.

no significant difference can be observed in the imposter rejection rate for different distances. This leads us to the conclusion that robust acoustic features, which suffer lesser variation with distance, are required to develop a speaker verification system robust for distant speech.

### 4. Short segment features derived from regions of high signal-to-noise ratio of speech

#### 4.1. Robustness of short segment features

Speech segments in the neighbourhood of the impulse-like excitations have a higher signal-to-noise ratio (SNR) because of the impulse-like nature of the excitation, when compared to other regions. It is likely that human beings extract information from high SNR regions in speech, which helps them perceive speech even at a distance [7]. Thus, features derived from short segments of speech signal in the regions of high SNR may be robust in case of distant speech.

We have used segments of 3 ms for computing the short-time spectrum, with a shift of 1 ms. Each frame of the speech signal is normalized so that the effect of noise on the dynamic range of spectrum is minimized. Figure 2(b) shows the wideband spectrogram of a speech signal captured from a close-speaking microphone. The sentence uttered is “We were away a year ago”. Notice that the formant contours are clearly visible in the spectrogram. To emphasize the formant contours, 10 successive short-time spectra are averaged over time with a shift of 5 frames. This results in a wideband spectrogram shown in Fig. 2(c).

Figure 3 shows the extraction of short-segment feature for the same signal collected at a distance of 6 ft. Notice the similarities between the wideband spectrograms of close-speaking and distant speech signals after the time-averaging of short-time spectra (Figs. 2(c) and 3(c)). The formant contours are clearly visible for close speech, and they degrade only slightly in the case of distant speech. This is not the case when spectral features are computed using 20-30 ms of speech, due to the inclusion of noisy samples in the computation of the short-time spectrum. Cepstral coefficients [8] are calculated from the time-averaged short-time spectra, to reduce the dimensionality.

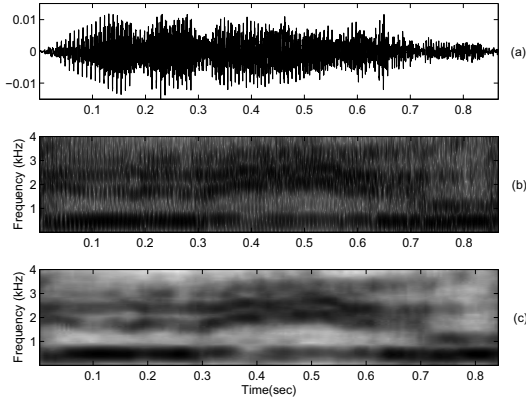


Figure 3: Robustness of short-segment analysis in distant speech. (a) Distant speech signal collected at 6ft, (b) wideband spectrogram of the speech signal, and (c) wideband spectrogram obtained after time-averaging the short-time spectra.

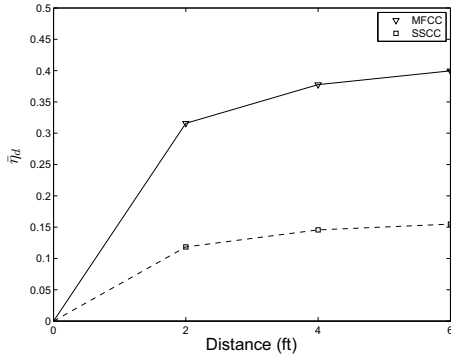


Figure 4: Variation of acoustic features with distance.

Twenty cepstral coefficients are derived to represent the spectral information in each frame. The extracted features are termed as short segment cepstral coefficients (SSCC).

We now examine the variation of MFCC and SSSC features with distance. The close-speaking speech signal and the corresponding distant speech signal are time-aligned, and MFCC and SSSC features are extracted from the corresponding frames of the signals. The Euclidean distance between the feature vectors  $\mathbf{x}_{c,i}$  and  $\mathbf{x}_{d,i}$ , derived from the  $i^{th}$  frame of close-speaking speech and distant speech respectively, is given by

$$\eta_d(i) = \|\mathbf{x}_{c,i} - \mathbf{x}_{d,i}\|. \quad (2)$$

where  $\mathbf{x}_{c,i}$  and  $\mathbf{x}_{d,i}$  are unit normalized vectors. The mean  $\bar{\eta}_d$  reflects the deviation of the feature as a function of distance. To compute  $\bar{\eta}_d$ , speech data from 50 utterances are used (10 speakers  $\times$  5 utterances/speaker). Only voiced frames are considered in the calculation of  $\bar{\eta}_d$ . The distances are computed for both MFCC and SSSC features. Figure 4 shows the average variation of the features across speakers as a function of distance. We observe from Fig. 4 that the average variation of SSSC features across distance is significantly less when compared to that of the MFCC features. This is attributed to the fact that SSCCs primarily represent the information specific to formants (spectral peaks), while MFCCs represent information specific to gross spectral envelope.

The speaker-specific nature of the SSCC feature is reflected in the performance of the modified baseline system, where the MFCC feature is replaced with SSCC feature. The second row of Table 1 shows that a significant improvement in the genuine acceptance rate is achieved for distant speech signals, by using the SSCC feature. A high imposter rejection rate is also obtained for all the distances.

#### 4.2. Significance of signal-to-reverberation component ratio in the selection of speech frames

Reverberation is the result of the addition of the reflected components of the signal to the direct component. The effect of the reflected components is dependent on the nature of speech segments. Hence all the segments may not suffer the same extent due to reverberation. In [9], the significance of regions of high signal-to-reverberant component ratio (SRR) was exploited for enhancement of reverberant speech. In this study, we use the normalized error of linear prediction (LP) analysis to detect the regions of high SRR in distant speech signal. The normalized error is the ratio of energy of the LP residual and that of the speech signal, computed over short segments (2-3 ms). The value of the normalized error is higher in the regions of high SRR, relative to other regions. An improvement in the genuine acceptance rate is observed, when the scores are computed using only the regions of high SRR (Table 1, row 3).

### 5. Significance of suprasegmental features for speaker verification

Human beings use several features like pitch, duration, speaking rate and speaking style for recognizing speakers. These features have complementary information, but have not been used extensively because of the difficulty in extraction and usage of those features. We exploit the duration and pitch of utterances for improving the performance of the speaker verification system.

#### 5.1. Duration

Duration information has been exploited for text-dependent speaker verification [10], where the nature of the optimal warping path obtained from the DTW algorithm is used to calculate the duration information. In this study, the DTW is performed using two sequences of SSSC features. The duration information is extracted by determining a line (regression line) that is best fit for the optimal warping path curve, and then measuring the deviation of the warping path from that line. Let  $(x_i, y_i)$ ,  $i = 1, 2, \dots, K$ , be the points on the warping path. The deviation of  $y_i$  from its regression line is an indication of the mismatch between two utterances. The regression line  $\hat{y}_i = mx_i + c$  can be found using the least squares method. The average sum of squared error ( $\xi$ ) represents the duration mismatch between the training and the verification features, and is given by

$$\xi = \frac{\sum_{i=1}^K (\hat{y}_i - y_i)^2}{K}. \quad (3)$$

The scores  $\alpha$  and  $\xi$  are normalized to the range of 0 to 1. A weighted sum  $S_1$  of the normalized  $\alpha$  and  $\xi$ , given by

$$S_1 = 0.9 \times \alpha + 0.1 \times \xi, \quad (4)$$

is used for speaker verification. The weights for the measures are empirically determined.

Table 1: Genuine acceptance rate of speaker verification system on close and distant speech.

Feature	close	2 ft	4 ft	6 ft
Baseline (MFCC + CMS)	91.7	43.2	22.6	20.5
SSCC	93.4	69.8	43.0	38.3
SSCC + high SRR	93.5	71.5	46.2	41.3
SSCC + duration	94.1	73.4	42.2	38.3
SSCC + pitch	95.1	72.6	46.3	42.2
SSCC + high SRR + duration + pitch	95.1	76.3	48.4	44.3

## 5.2. Pitch

Though pitch is a robust feature in distant speech, difficulty exists in the extraction of fundamental frequency ( $F_0$ ) from distant speech signals. A robust method was proposed in [11] for extraction of  $F_0$  from clean and noisy speech signals based on the strength of impulse-like excitations in voiced speech. We use the method for extracting  $F_0$  from distant speech signals. The similarity of the pitch contours of the reference and test utterances is captured using the optimal warping path obtained from the DTW algorithm [10], where the absolute difference of the  $F_0$  values for the selected matching frames in the reference and test utterances are summed up to obtain the pitch score  $\tau$ , given by

$$\tau = \sum_{i=1}^L |F_0(x_i) - F_0(y_i)|. \quad (5)$$

Here  $F_0(x_i)$  is the fundamental frequency of the frame  $x_i$  of the test utterance,  $F_0(y_i)$  is the fundamental frequency of the frame  $y_i$  of the reference utterance and  $L$  is the number of points (20) chosen for computing the score. These points correspond to the least Euclidean distance in the optimal warping path. Similar to the duration information, a weighted sum  $S_2$  of the normalized  $\alpha$  and  $\tau$  is used in the final score computation, which is given by

$$S_2 = 0.9 \times \alpha + 0.1 \times \tau. \quad (6)$$

The use of duration and pitch information does help in improving the performance of the text-dependent speaker verification system, as observed from an improved genuine acceptance rate (Table 1, rows 4 and 5).

## 5.3. Combination of features

A study is made on the combination of SSCC features extracted from high SRR regions, duration and pitch information. For this purpose, all the scores are normalized to the range of 0 to 1, and a weighted sum  $S_3$  of the measures is calculated as

$$S_3 = 0.8 \times \alpha + 0.1 \times \xi + 0.1 \times \tau. \quad (7)$$

The weights are determined empirically. The additional information of high SRR regions, duration and pitch was able to improve the performance of the system for distant speech as shown in Tables 1 and 2. Improvements can be seen both in the genuine acceptance and imposter rejection rates.

## 6. Conclusions

This paper attempts to exploit the robustness of subsegmental and suprasegmental features of speech, for the task of speaker verification using distant speech signals. An acoustic feature

Table 2: Imposter rejection rate of speaker verification system on close and distant speech.

Feature	close	2 ft	4 ft	6 ft
Baseline (MFCC + CMS)	99.3	99.7	99.8	99.8
SSCC	99.3	99.5	99.7	99.6
SSCC + high SRR	99.4	99.5	99.5	99.4
SSCC + duration	99.5	99.7	99.8	99.8
SSCC + pitch	99.4	99.7	99.8	99.7
SSCC + high SRR + duration + pitch	99.5	99.8	99.8	99.8

derived from short segments of speech signal was proposed. The proposed feature (SSCC) exploits the high signal-to-noise nature of speech signal in the neighbourhood of significant excitations using short segments. It was shown that the proposed feature suffers lesser degradation with distance when compared to the MFCC feature. The feature also yields a better performance of speaker verification for distant speech compared to the MFCCs. This is attributed to the fact that SSCCs primarily represent the information specific to the formants (spectral peaks). The performance of the system using the SSCC features was enhanced by considering the high SRR regions for score computation. Additional information of duration and pitch was used to further improve the performance of the speaker verification system.

## 7. References

- [1] G. Doddington, "Speaker recognition-identifying people by their voices", Proc. IEEE, vol. 73, no. 11, pp. 1651–1664, Nov. 1985.
- [2] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition", IEEE Trans. Audio, Speech Lang. Process., vol. 15, no. 7, pp. 2023–2032, Sep. 2007.
- [3] I. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays", in Proc. A Speaker Odyssey - the Speaker recognition Workshop, Crete, Greece, pp. 101–106, June 2001.
- [4] J. Gammal and R. Goubran, "Combating reverberation in speaker verification", in Proc. IEEE Instrumentation and Measurement Technology Conference, Canada, vol. 1, pp. 687–690, May 2005.
- [5] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays", in Proc. Int. Conf. Spoken Language Processing, Philadelphia, PA, USA, vol. 3, pp. 1333–1336, Oct. 1996.
- [6] S. Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, no. 2, pp. 254–272, Apr. 1981.
- [7] O. Ghitza, "Robustness against noise : The role of timing-synchrony measurement", in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing, Dallas, Texas, USA, vol. 4, pp. 2372–2375, Apr. 1987.
- [8] L. Rabiner and B. H. Juang, Fundamentals of speech recognition. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
- [9] B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendano, and H. Hermansky, "Enhancement of reverberant speech using LP residual", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 1, pp. 405–408, May 1998.
- [10] B. Yegnanarayana, S. Prasanna, J. Zachariah, and C. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system", IEEE Trans. Speech Audio Processing, vol. 13, no. 4, pp. 575–582, Jul. 2005.
- [11] B. Yegnanarayana and K. S. R. Murty, "Event-Based Instantaneous Fundamental Frequency Estimation From Speech Signals", IEEE Trans. Audio, Speech Lang. Process., vol. 17, no. 4, pp. 614–624, May 2009.