Throat Microphone Signal for Speaker Recognition

A. Shahina and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engg.
Indian Institute of Technology Madras,
Chennai - 600 036, India.
{shahina,yegna}@cs.iitm.ernet.in

M. R. Kesheorey

Scientist
Center for Artificial Intelligence and Robotics
Bangalore 560 001, India.
cair3@vsnl.net

Abstract

Speaker recognition systems perform better when clean speech signals are used for the task. In the presence of high levels of background noise, speech recorded from a close speaking microphone will be degraded and hence the performance of the speaker recognition system. Use of a transducer held at the throat results in a signal that is clean even in a noisy environment. This paper discusses the prospect of using such signals for speaker recognition. A study of a text-independent speaker recognition system based on features extracted from speech simultaneously recorded using a throat microphone and a close-speaking microphone in clean and simulated noisy conditions is conducted. Autoassociative neural networks are used to model the speaker characteristics based on the vocal tract system and excitation source features represented by weighted linear prediction cepstral coefficients and linear prediction residual, respectively. The results of experimental studies show that the speech collected from the throat microphone can be used for tasks like speaker recognition, especially in noisy conditions.

1. Introduction

A throat microphone, placed in contact with the skin surrounding the larynx near the vocal folds, picks up its vibrations and also the signals transmitted through the muscles of the speech production mechanism. The resulting signal (hereafter called throat speech) is very similar to normal speech. Due to its proximity to the speech production system, speech recorded from a throat microphone is clean, and is not affected by environmental noise. The intelligibility of the throat speech signal is almost equivalent to that of the speech signal recorded using the microphone placed close to the speaker (hereafter called close-speaking). In a noisy environment, the intelligibility of a close-speaking microphone speech is affected, as the microphone picks up not only the voice but also the background noise and reflections from various objects. But the intelligibility of the throat microphone signal is nearly the same as that of the signal obtained in a noise-free environment. Hence the throat microphone is a preferred choice for use in speech applications even in adverse conditions.

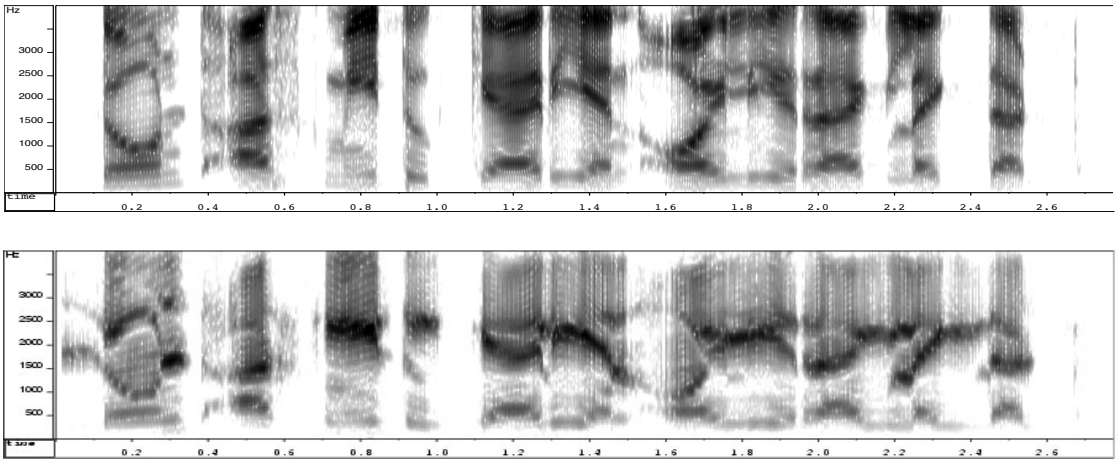
Applications such as entry into high-security enclosures and access control may involve noisy environments, for instance, cockpit of an aircraft. For such applications, a reliable person identification is required. Speaker recognition is the task of person identification using speech as the biometric feature [1][2]. A person's voice, like other biometrics (finger prints, retinal patterns or genetic structure), cannot be forgotten or misplaced unlike the use of artifacts for identification by artificial means such as keys or memorized passwords [3][4]. Hence, speaker recognition is more reliable than other artifacts for person identification. As mentioned earlier, in adverse conditions, a person's voice is less affected when recorded using a throat microphone than when using a close-speaking microphone. The objective of this study is to analyse the characteristics of the speech collected using the throat microphone, and illustrate the presence of significant speaker-specific information in the vocal tract system and excitation source characteristics. It is also interesting to note that for applications other than synthesis, we need features that are robust against degradations, rather than features that are useful for intelligibility and quality.

This paper is organized as follows: In Section 2, we analyze the characteristics of a throat microphone speech by comparing with that of a close-speaking microphone speech. Section 3 discusses the speaker recognition studies conducted using the speech collected simultaneously from the throat and close-speaking microphones. Autoassociative neural network (AANN) models are used for capturing the speaker-specific information present in the signals. Section 4 concludes with the various issues discussed in this study.

2. Analysis of throat microphone speech signal

The throat microphone is a transducer that is placed in contact with the skin surrounding the larynx near the vocal folds. The throat microphone converts the vibrations that it picks up into equivalent speech signals. Typically, the throat speech is a low amplitude signal when compared to the close-speaking microphone signal. But, it is interesting to note that the throat speech is of high quality. The throat microphone speech is relatively unaffected by background noise and reverberation effects. This signal is almost as intelligible as the close-speaking microphone speech. Figure 1 shows the wideband spectrograms of the speech from a male speaker recorded simultaneously using the throat and the close-speaking microphones for the sentence *don't ask me to carry an oily rag like that*. The wideband spectrogram of the throat speech shows that the first three formants are present as in the close-speaking microphone speech, but unlike the later, the higher formants are not well represented (refer Figures.1(a) and 1(b)).

As it can be seen from the spectrograms, voiced stop consonants like /d/ and /g/ are represented better in the case of throat speech. In contrast, nasal consonants like /m/ are poorly represented in the throat speech. Although there are a few differences between the two signals as mentioned above, the overall intelligibility of the speech signals are nearly the same.



(a)

(b)

Figure 1: The wide band spectrograms of a speech signal from a male speaker recorded simultaneously from a close speaking microphone and a throat microphone ((a) and (b) respectively) for the sentence *don't ask me to carry an oily rag like that*.

In order to compare the features of the speech collected from the two microphones, simultaneous acquisition of speech using the throat and close-speaking microphones was carried out. The signals were sampled at 8 kHz. In this study, features pertaining to the vocal tract system and the excitation source characteristics are used. To obtain these features, a 12^{th} order LP analysis is performed on overlapping Hamming windowed speech frames of 20 msec duration taken with a frame shift of 5 msec duration [5]. In the LP analysis, the n^{th} speech sample is predicted as a linear weighted sum of the past p samples, and is given by:

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

where $\{a_k\}$ are the model parameters and p is the order of prediction. The difference between the actual value $s(n)$ and the predicted value $\hat{s}(n)$ of the n^{th} speech sample is the prediction error, known as the LP residual and given by:

$$e(n) = s(n) - \hat{s}(n) \tag{2}$$

The LP coefficients (representing the vocal tract characteristics) are converted to 19 linearly weighted LP cepstral coefficients (WLPCC) [6]. The cepstral coefficients are obtained using the following relations:

$$c_o = \ln \sigma^2 \tag{3}$$

$$c_m = -a_m - \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \quad 1 \leq \mathrm{m} \leq \mathrm{p} \tag{4}$$

$$c_m = -\sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \quad \mathrm{m} > \mathrm{p} \tag{5}$$

where σ^2 is the gain term in the LPC model, c_m is the LPCC, and a_m is the LP coefficient. These 19 linearly weighted cepstral coefficients (WLPCCs) corresponding to each speech frame are used as the system features in this study. The LP residual obtained after removing the vocal tract system features mostly contains information about the excitation source. Hence, the LP residual is used for extracting speaker-specific source features for this study.

3. Speaker recognition studies

3.1. Database for the study

Recording was done in the laboratory under clean and noisy conditions. Noisy environment was simulated using radio static. Speech from volunteers was acquired simultaneously using the throat and close-speaking microphones. Text-independent speech was used in this study. Two minutes of speech data obtained from each of the 40 speakers is used to train a speaker model. Each test utterance was of 20 secs duration. The recordings for training and testing the speaker models were carried out in separate sessions. About 240 test utterances obtained from the 40 speakers under clean and noisy conditions were used in this study.

3.2. AANN models for speaker recognition

The feature vectors representing the speech data have a complex distribution in the multi-dimensional feature space, and the surface representing this distribution may be highly nonlinear. The potential of artificial neural networks as nonlinear models is exploited to capture the characteristics of the vectors unique to a speaker from the given training data [7][8]. Specifically, the autoassociative neural network models, which are feedforward neural networks that perform the task of autoassociation are used. It has been shown that AANN models capture the distribution of the feature vectors in the high dimensional space. The training error surface relates to the distribution of the given feature vectors [9]. Typical structure of a five layer AANN used in this study is shown in Figure 2.

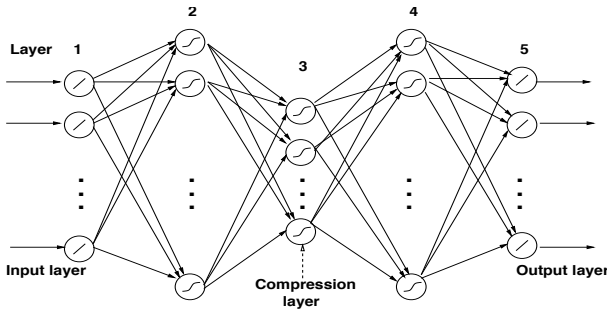


Figure 2: Five layer AANN model.

3.3. Speaker recognition using system features

The structure of the neural network used is *19L 38N 4N 38N 19L* where L refers to a linear unit and N to a nonlinear unit, the numbers represent the number of nodes in a layer. The 19-dimension system features (WLPCC) obtained for each speaker are given to the AANN in a randomized fashion. Two separate speaker models are obtained by training two AANN models using WLPCCs derived from clean and noisy speech. Each AANN is trained using 200 epochs.

3.4. Speaker recognition using source features

To implicitly learn the speaker information present in the excitation source, LP residual down-sampled to 4 kHz sampling frequency to emphasize only those regions with high signal-to-noise ratio are used. Blocks of 20 samples (5 msec) of the normalized LP residual, with a shift of one sample, are applied in succession. The structure of the AANN used is *20L 40N 10N 40N 20L*. Two speaker models are obtained for each speaker using the LP residual derived from clean and noisy speech. A model is trained using 200 epochs.

3.5. Testing of system and source models

Test utterances from clean and noisy environments, each of 20 secs duration, were used to test the source and system based speaker models trained using clean speech. The noisy test utterances were also used to test speaker models trained using noisy speech. Both the source and system features are extracted using a 12^{th} order LP analysis. The 19 dimension WLPCCs and blocks of 20 samples of the LP residual shifted by one sample form the input to the system and source models, respectively [10]. The deviation of the output of each model from its input is used to compute the squared error E_i for each frame or block. This error is used to compute the confidence score of that frame or block, which is used as a performance measure for the speaker recognition system. The confidence score C_i of the i^{th} frame or block is expressed as $C_i = exp(-\lambda E_i)$ where the constant λ is set to 1 in this study. This confidence value is higher if the error is lower, when the frame or block of the test utterance of a speaker matches with the corresponding model. When the frame or block of the test utterance does not match with the corresponding model, the error is high, and this lowers the confidence score of that frame or block. A test utterance is compared with each speaker model to obtain the average confidence score C, which is expressed as $C = 1/N \sum_{i=1}^{N} C_i$, where N is the total number of frames/blocks. The average confidence scores of the test utterance against all the models are compared to evaluate the performance of the speaker recognition systems based on the system and source features derived from clean and noisy speech.

3.6. Performance evaluation

Performance of the speaker recognition systems based on the system and source features derived from the clean speech obtained from throat and close-speaking microphones is given in Table 1. The performance is evaluated in terms of percentage of the number of test utterances accepted out of the total test utterances used for this study. We see that the performance of the speaker recognition system using the throat microphone speech is similar to that using the close-speaking microphone speech for both system as well as source features. As the scores obtained for both the system and source feature based models are from independent evidences, the two scores can be combined. The combination logic is a simple addition of the two scores. The block diagram of the proposed speaker recognition system using the combined evidences is shown in Figure. 3. The speaker recognition system based on combined scores performed relatively better in the case of throat microphone speech than for the close-speaking microphone speech.

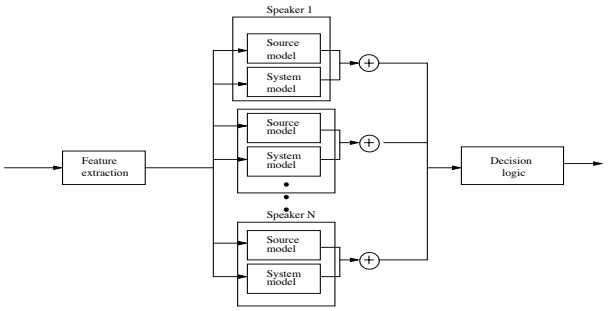


Figure 3: Block diagram of the speaker recognition system using combined evidences.

Table 2 shows the performance of the speaker recognition systems, where the speaker models trained using clean speech are tested against noisy utterances. The performance of the speaker recognition systems based on both the system and source features is very poor in the case of the close-speaking microphone. Though the throat microphone based speaker recognition systems perform relatively better, this performance is poor compared to the recognition systems using clean utterances. Since the throat microphone is relatively immune to ambient noise the drop in performance can be attributed to the significant change in the speaker characteristics. Speakers tend to stress the syllables and speak aloud in the presence of noise (Lombard effect). The poor performance of the close-speaking microphone based speaker recognition systems is due, both to the change in the speaker characteristics and the presence of significant noise levels in the speech.

In order to verify that the degradation in performance of the throat microphone based system is primarily due to the change in speaker characteristics, we used AANN models trained using noisy speech. These models were tested against noisy speech. This ensures that the voice characteristics of the speaker is similar in the training and testing stage. We observe that the performance of the throat microphone based speaker recognition systems is similar to that of clean speech based speaker recognition systems (refer Table 3). But, the performance of the close-speaking microphone based speaker recognition systems

Performance (%) of the speaker recognition systems based on source and system features obtained from simultaneously recorded speech signals using throat and close speaking microphones

Table 1: Models trained and tested using speech in clean environment

Speech	System features	Source features	Combined system
Throat microphone	84.3	73.0	94.3
Close-speaking microphone	84.3	70.0	88.6

Table 2: Models trained using speech in clean environment and tested in noisy conditions

Speech	System features	Source features	Combined system
Throat microphone	50.0	27.0	50.0
Close-speaking microphone	8.3	10.0	16.67

Table 3: Models trained and tested using speech in noisy environment

Speech	System features	Source features	Combined system
Throat microphone	83.3	75.0	93.3
Close-speaking microphone	19.42	25.0	25.0

is poor. This is due to degradation in the speech.

4. Summary and conclusions

In this paper, we have analyzed the characteristics of the speech signals collected from a throat microphone. From the studies conducted, we infer that the throat microphone speech indeed contains significant information about vocal tract system and excitation source characteristics. The performance of the recognition systems using throat and close-speaking microphone speech signals, when the signals are recorded under noise-free conditions are almost similar. The performance of the systems based on both the throat microphone speech and close-speaking microphone speech degrades when utterances recorded under noisy conditions are used to test speaker models trained using clean speech signals. The degradation in performance in the former is due to the Lombard effect, whereas the degradation in performance in the later is due to both the Lombard effect and background noise. When the Lombard effect is minimized by training and testing using speech recorded under noisy conditions we observe that the performance of the throat microphone based systems is similar to that under noise-free conditions. However, the performance of the close-speaking microphone based recognition systems do not improve significantly. This shows that the performance of the system using close-speaking microphone data degrades as the background noise increases, whereas the performance of the system using throat microphone data is likely to be unaffected by the background noise.

5. References

[1] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.

[2] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine*, vol. 3, pp. 4–17, Oct. 1986.

[3] J. P. Campbell, Jr, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sept. 1997.

[4] G. R. Doddington, "Speaker recognition-identifying people by their voices," *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, Nov. 1985.

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[6] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[7] B. Yegnanarayana, *Artificial Neural Networks*, Prentice Hall India, Connaught Circle, New Delhi, 1999.

[8] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall International, NJ, 1999.

[9] B. Yegnanarayana and S. P. Kishore, "AANN: An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, no. 3, pp. 459–469, Apr. 2002.

[10] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, May, pp. 409–412.