Significance of Formants from Difference Spectrum for Speaker Identification

Kishore Prahallad[*+], *Sudhakar Varanasi*, *Ranganatham Veluru*, *Bharat Krishna M*, *Debashish S Roy*

*Language Technologies Institute, Carnegie Mellon University
+International Institute of Information Technology, Hyderabad, India
skishore@cs.cmu.edu

Emergency Response System, Satyam Computer Services Limited
sudhakar_varanasi@satyam.com, ranganatham_veluru@satyam.com, bharat_masimukku@satyam.com, debashish_swapankumar@satyam.com

Abstract

In this paper, we describe a prototype speaker identification system using auto-associative neural network (AANN) and formant features. Our experiments demonstrate that formants extracted from difference spectrum perform significantly better than formants extracted from normal spectrum for the task of speaker identification. We also demonstrate that formants from difference spectrum provide comparable speaker identification performance with that of features such as weighted linear predictive Cepstral coefficients and Mel-Frequency Cepstral coefficients. Finally, we combine the results of formant based system and linear predictive Cepstral coefficients based system to achieve 100% identification performance.

Index Terms: Formant extraction, difference spectrum, speaker identification, autoassociative neural network models

1. Introduction

An open set speaker identification system attributes the identity of a test sample to one of the registered speakers or declares that it does not belong to any of the registered speakers. In order to build a speaker identification system, we investigated the use of auto associative neural network (AANN) models and experimented with several features including formants extracted from a normal spectrum as well as from the difference spectrum (ref. to Sec 4 for details). On a database of 50 speakers our experiments showed that formants extracted from difference spectrum perform significantly better than the formants from normal spectrum. Our experiments also demonstrated that the performance of formants based identification system is comparable to that of weighted linear predictive Cepstral coefficients (WLPCC). Moreover, the combining of results from formant based identification system and WLPCC based system provided 100% identification.

This paper is organized as follows: Section 2 describes the AANN models and their distribution capturing ability. Section 3 discusses the development of speaker identification system using AANN models and the base line results using WLPCC and Mel-frequency Cepstral Coefficients (MFCC). Section 4 describes the difference spectrum and extraction of formants. Section 4.2 discusses the results using formants and compares its performance with that of WLPCC and MFCC.

2. Auto Associative Neural Networks

AANN models are feed forward neural networks performing identity mapping of the input space [1][2]. The network architecture of these models may have more than one hidden layer [3]. The input layer and the output layer have same number of processing units. One of the hidden layers known as the bottleneck layer or dimension compression layer has smaller dimension than the input layer. These networks can be trained using algorithms such as back propagation to reconstruct the input data at the output layer. The units of the dimension compression hidden layer represent the significant features of the input data like in the case of principal component analysis. This characteristic of AANN model is exploited extensively for linear and nonlinear dimension compression of the input data [3][4]. Studies reported in [5][6][7] have demonstrated that the AANN models can also be used as nonparametric models to capture the distribution of the input data. It has also been demonstrated in [5][6][7] that speaker models could be built using AANN for the task of speaker verification. In this paper, we exploit the distribution capturing ability of AANN models to build a speaker identification system.

3. AANN Based Speaker Identification

In this work, we have developed a speaker identification system using a database of 50 speakers. These are an arbitrary set of 50 speakers belonging to different ethnic groups speaking different languages and different sections of society. Each speaker has an AANN model, and our speaker identification system has three phases: (1) Speaker Enrollment (2) Speaker Training and (3) Speaker Testing (Identification)

3.1. Enrollment Phase

During enrollment, the speaker reads an arbitrary text of any language for about three minutes. Speech signal is recorded in laboratory environment using the recorder facilities provided by the computer system. The recorded signal is sampled at 16000 Hz. The three minutes of recorded speech in split into one training wave file of length 84 sec and two testing wave files each of length 48 sec.

3.2. Training Phase

During training an AANN model is built for each speaker. The structure of the AANN model used is given by 17L39N10N39N17L in the case of WLPCC and 13L39N10N39N13L in the case of MFCC, where L denotes linear and N denotes non linear units. The integer value denotes the number of layers in that particular area. The output generated by the training phase consists of weight files corresponding to all the synaptic connections between neurons of adjacent layers.

3.3. Testing Phase

During testing the hypothesis is that if the test wave-file from a speaker is given to an AANN model of the same speaker then the error would be less compared to the error generated in the case when the AANN model and test wave-files do not belong to same speaker.

The testing process could be explained as a sequence of below steps.

- Let X(i) denote the set of input vectors of the test-wave file, where i is the index of the feature vector, and M(k) is the AANN model of speaker k.
- For each speaker k in the database
  - Give X as input to AANN model M(k) and obtain output vectors say Y(k).
  - Compute the average Euclidean distance ED(k) between X and Y(k).
- Let $\min_k \arg\{ED(k)\}$ = j, where j is the index of the speaker model giving the minimum Euclidean.
- If j also happens to be the actual speaker of the test wave-file, then it can be said that the correct speaker has been identified.

3.4. Baseline System Results

To build the baseline system, feature vectors were derived from the speech signal using a frame size of 10 ms and a frame shift of 5 ms. A 12^{th} linear prediction was performed on the speech signal to derive 17 Cepstral coefficients which were linearly weighted to form 17-dimensional WLPCC. On each frame Mel-frequency filters were also employed to obtain 13 dimensional MFCC features. The performance of the speaker identification system on a subset of 25 speakers with 50 test-samples (25 speakers x 2 test-samples) and on the full set of 50 speakers with 100 test-samples (50 speakers x 2 test-samples) using MFCC and WLPCC is shown in Table 1. Experiments were conducted on the subset of 25 speakers to tune the structure of AANN model specifically in the case of formants.

Table 1: *Base Line System performance with WLPCC and MFCC features*

Exp No	Feature	Speakers Identified out of 50	Performance (%)	AANN Structure
1	MFCC	41/50	82	13L39N10N39N13L
2	WLPCC	48/50	**96**	17L39N10N39N17L
3	MFCC	57/100	57	13L39N10N39N13L
4	WLPCC	97/100	**97**	17L39N10N39N17L

It can be observed from Table 1 that Weighted LPCC performs better than MFCC on this small set of speakers. In order to investigate the formants as features for speaker identification and compare their performance with that of WLPCC and MFCC we conducted a series of experiments which are described in the following sections.

4. Extraction of Formants

Formants are the resonances of the vocal tract and are identified with the peaks in short-time Fourier transform. These peaks can be extracted either from the smoothed Fast Fourier Transform (FFT) or from the Linear Prediction (LP) spectrum [9]. In model based techniques such as LP, the number of peaks that could be present in the spectrum is dictated by the order of the linear prediction [10]. A lower order (4-6) linear prediction models the spectrum with the dominant peaks, whereas, a higher order (10-16) would approximate most of the peaks and is a better representation of the spectrum.

Typically a peak picking algorithm is employed to pick peaks from a higher order LP spectrum (referred to as normal spectrum here). In [8], it is proposed that the difference spectrum could be used for better estimation of formants. In this paper we use the formants extracted from difference spectrum for the task of speaker identification.

4.1. Formants from Difference Spectrum

In LP analysis of speech, an all pole model is assumed for the system producing the signal s(n). The power spectrum of the sampled signal is modeled in an optimal manner by an all pole model spectrum. Let us assume that the model spectrum corresponds to a transfer function $H(z)$ given by

$$H(z) = \frac{G}{A(z)} = \frac{G}{1+\sum_{k=1}^{p} z^{-k}}$$ Where $A(z)$ is called inverse filter and p is the number of poles in the model spectrum, and G is the gain factor. The model spectrum $P'(z)$ is given by $P'(z) = |H(z)|^2$. Let $P'(z)$ and

$P''(z)$ represent the all-pole spectrums of $s(n)$ obtained for two different orders p, q of LP analysis respectively. Let $Q(z)$ represent the difference spectrum of $P'(z)$ and $P''(z)$ and is given by

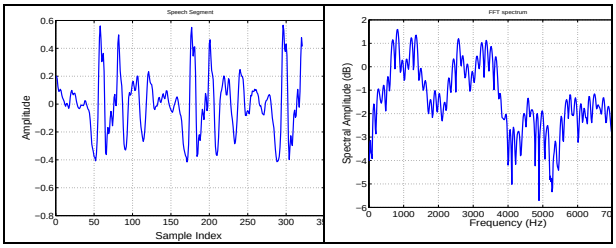
$$Q(z) = \log P''(z) - \log P'(z)$$

$$Q(z) = \log \frac{P''(z)}{P'(z)}$$

The difference spectrum $Q(z)$ in the logarithm domain is the log ratio of $P''(z)$ and $P'(z)$. When $p < q$ then $P'(z)$ represents the lower order LP spectrum and $P''(z)$ represents the higher order LP spectrum of $s(n)$. If $P''(z) < P'(z)$ then the difference spectrum is negative and it de-emphasizes the regions which are lower values in $P''(z)$ and the corresponding regions are higher values in $P'(z)$. If $P''(z) > P'(z)$ then the difference spectrum is positive and these regions of $P''(z)$ get boosted up in the difference spectrum.

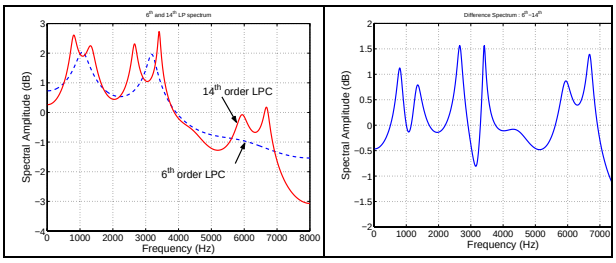
Fig. 1(a) shows a short speech segment s(n) and Fig. 1(b) shows the spectrum of s(n) computed using FFT. Fig. 2(a) shows the 6th and 14th order LP spectrums of s(n) and Fig. 2(b) shows their difference spectrum. It can be observed that the peaks of 14th order LP spectrum are boosted in the difference spectrum.

Another interesting thing to observe is the difference spectrum of 14th and 13th order LP spectrums. Fig. 3(a) shows the 13th and 14th order LP spectrums of s(n) and Fig. 3(b) shows their difference spectrum. The peaks are better emphasized in the difference spectrum obtained from 14th and 13th order LP difference spectrum. It also picks up the peak at 4400 Hz, which is not dominantly represented in the 14th order LP spectrum.



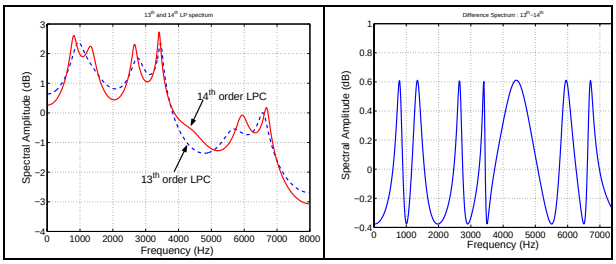
(a) **(b)**

Figure 1: *(a) Speech segment s(n) (b) FFT of s(n)*



(a) **(b)**

Figure 2: (a) 6th and 14th order LP spectrum of a signal s(n) (b) Difference Spectrum of 14th and 6th order LP spectrums.



(a) **(b)**

Figure 3: (a) 13th and 14th order LP spectrum of a signal s(n) (b) Difference Spectrum of 14th and 13th order LP spectrums.

It is conjectured that difference operation on $P''(z)$ and $P'(z)$ in log domain results in removing the varying slope (damping effect due to glottal roll off) thus emphasizing the peaks of $P''(z)$ as observed in Fig. 2(b) and Fig. 3(b). Such an emphasis would aid peak picking algorithms to locate the formants easily.

4.2. Speaker Identification using Formants

Let F_DS denote formants extracted from difference spectrum and F_S denote formants extracted from normal spectrum. The front end of the base line system was modified to extract F_S and F_DS from the speech signal. Variance normalization was performed on these features, and speaker models were built by changing the structure of the AANN. The performance of speaker identification system for F_S and F_DS is as shown in Table 2. The "#" column in Table 2 indicates that the number of formants that were extracted for each frame. It was found that 6 or 8 formants performed better than that of 4 formants for speaker identification. This observation matches with the intuition that the higher formants carry speaker information. Use of higher formants denotes more details about vocal tract shape which are needed for the task of speaker identification. From experiments 4,5 and 6,7 and 8,9 in Table 2, it can be observed that F_DS performs significantly better than that of F_S for speaker identification. These results provide empirical evidence that useful features could be extracted from F_DS. From Table 1 and Table 2, it could also be observed that an identification performance of 92% was achieved by F_DS in comparison with 97% by WLPCC.

Table 2: *Speaker Identification System performance using formants. Experiments 8 and 9 were performed on 50 speakers' database, while the rest of them were performed on a subset of 25 speakers.*

Exp No.	Formants		Speakers Identified	Performance (%)	AANN Structure
	#	Type			
1	4	F_S	2/50	04	4L39N4N39N4L
2	4	F_DS	6/50	12	4L39N4N39N4L
3	4	F_DS	21/50	42	4L39N2N39N4L
4	6	F_S	06/50	12	6L39N4N39N6L
5	6	F_DS	46/50	**92**	6L39N4N39N6L
6	8	F_S	07/50	14	8L39N4N39N8L
7	8	F_DS	46/50	**92**	8L39N4N39N8L
8	6	F_S	07/100	07	6L39N4N39N6L
9	6	F_DS	92/100	**92**	6L39N4N39N6L

It was also found that a set of speakers (say X) were not identified by WLPCC based speaker identification system, but were identified by F_DS based speaker identification system. These speaker models performed better in F_DS feature domain than in WLPCC feature domain. An identification performance of 100% was achieved when we used F_DS based AANN models for the speakers belonging to set X, and WLPCC based AANN models for the rest of the speakers during identification.

In order to investigate further on the role of higher formants for the task of speaker identification, we considered only the last four formants out of six formants extracted from the speech signal. The performance of speaker identification system obtained using the last four formants is as shown in Table 3. The lower performance of this identification system in comparison with 92% obtained for experiment 5 in Table 2 indicates that the first two formants do carry significant information about the speaker and are important for the task of identifying the speaker.

Table 3: ***Performance of speaker identification system using last four of six formants. The number of dimension compression units in the AANN structure is being changed in these two experiments.***

Formants	Type	Speakers Identified	Performance in %	AANN Structure
last 4 out of 6	F_DS	24/50	48	4L39N3N 39N4L
Last 4 out of 6	F_DS	29/50	58	4L39N2N39N4L

5. Conclusions

In this paper, we have discussed the prototype system built for the task of speaker identification. This system uses distribution capturing ability of AANN models for the task of speaker identification. While the deployable system has to be tuned to telephone or clean speech for a much larger dataset, our motivation of building the prototype system of 50 speakers database of microphone speech was to experiment with different features, and tune the neural network architecture to the possible extent.

The experiments using formant features indicated that the formants extracted from difference spectrum performed significantly better than that of formants extracted from normal spectrum for the task of speaker identification. These results provided empirical evidence that useful features could be extracted from difference spectrum, and the formants extracted from difference spectrum are better estimated than that of from normal spectrum. It was also shown that the performance of formants from difference spectrum was comparable to that of weighted linear prediction Cepstral coefficients. Combining evidences from formant based system and linear predictive coefficients based system provided 100% identification results on a dataset of 50 speakers. Our future work lies in using these formants for robust speaker identification on telephone speech on a much larger corpus. We are also working to develop the open set speaker identification by adapting some of the verification techniques.

6. Acknowledgements

We are grateful to Prof. Raj Reddy, Carnegie Mellon University for promoting the collaboration between ERS Team of Satyam Computer Services Ltd. and IIIT Hyderabad and making this collaborative work possible. We would like to thank Rohit Kumar, Carnegie Mellon University for his valuable comments on the draft. We would also like to thank Anil Varma Sayyaparaju and Sri Lakshmi B for their active contribution in the initial stages of the project. We are also thankful to Bharadwaj R, Humera N, Ram Chander PSS and Paramita C for doing the code review and code walkthrough. We also thank Liz Mary James, Priyanka S and Venkateswaramma B for their contribution in documenting and also in implementing the CMMI level 5 quality processes for this project.

7. References

[1] B. Yegnanarayana, *Artificial Neural Networks.* New Delhi: Prentice Hall of India, 1999.

[2] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New Jersey: Prentice Hall Inc., 1999.

[3] M. A. Kramer, "Nonlinear Principle Component Analysis using Autoassociative Neural Networks," *AIChE,* Vol. 37, pp. 233 – 243, Feb. 1991.

[4] K. I. Diamantaras and S. Y. Kung, *Principle Component Neural Networks: Theory and Applications.* New York: John Wiley & Sons Inc., 1996.

[5] B. Yegnanarayana and S. P. Kishore, "AANN – an alternative to GMM Pattern Recognition," *Neural Networks Communicated.*

[6] M. Shajith Ikbal, *Autoassociative Neural Network Models for Speaker Verification.* MS Thesis, Dept. of CSE, IIT Madras, 1999.

[7] S. P. Kishore, *Speaker Verification Using Autoassociative Neural Network Models for Speaker Verification.* MS Thesis, Dept. of CSE, IIT Madras, 2001.

[8] Kishore Prahallad, Vamshi Ambati and B. Yegnanarayana, *Formant Extraction Using Difference Spectrum.* Unpublished Technical Report 2004.

[9] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, 1993.

[10] J. Makhoul, *Linear Prediction: A Tutorial Review.* Proc. IEEE, vol. 63, no. 4, pp. 561-580, Apr. 1975.