Significance of Pitch Synchronous Analysis for Speaker Recognition using AANN Models

Sri Harish Reddy M, Kishore Prahallad, Suryakanth V. Gangashetty, and B. Yegnanarayana

International Institute of Information Technology, Hyderabad, India
sriharsham@research.iiit.ac.in, {kishore,svg,yegna}@iiit.ac.in

Abstract

For speaker recognition studies, it is necessary to process the speech signal suitably to capture the speaker-specific information. There is complementary speaker-specific information in the excitation source and vocal tract system characteristics. Therefore it is necessary to separate these components, even approximately, from the speech signal. We propose linear prediction (LP) residual and LP coefficients to represent these two components. Analysis is performed in a pitch synchronous manner in order to focus on the significant portion of the speech signal in each glottal cycle, and also to reduce the artifacts of digital signal processing on the extracted features. Finally, the speaker-specific information is captured from the excitation and the vocal tract system components using autoassociative neural networks (AANN) models. We show that the pitch synchronous extraction of information from the residual and vocal tract system bring out the speaker-specific information much better than using the pitch asynchronous analysis as in the traditional block processing using an analysis window of fixed size.

Index Terms: Speaker recognition, pitch synchronous, AANN, glottal closure instants.

1. Introduction

One of the main issues in speaker recognition task is to extract features specific to a given speaker. If possible, these features should be captured from a small amount of data during both training and testing phases. Also, the features should be robust against degradation in speech due to channel and noise. Over the past several years, many attempts have been made to capture the speaker-specific information in a model from a large amount of training data, and test the model using a limited amount of data. The search for features spans over several dimensions, such as at language level, nonverbal gestures, suprasegmental ($>$ 100 ms) features, segmental (10-30 ms) features and subsegmental (1-3 ms) features. At the language level, one can use an automatic speech recognition (ASR) system to determine the usage of a subset of word combinations specific to a speaker. This requires a robust ASR, and also a large amount of data from each speaker. Nonverbal gestures include use of non-speech acoustic sounds like $umm's$, $ah's$ and with some other user habits such as 'you see', 'I mean' etc. Identifying and detecting these gestures is a challenging task. Also, a large amount of data is needed from each speaker in order to get sufficient examples to identify a speaker. Features at the suprasegmental level are usually prosody features, consisting of intonation and duration patterns of a given speaker. The suprasegmental features are not only specific to a language and an environment, but they are also acquired by an individual over a period of time. The segmental features correspond mostly to the acoustic characteristics of individual sound units, and they reflect the vocal tract system and source characteristics in a short-segment of over 10-50 ms of speech. The segmental features are typically the short-term spectral features reflecting the vocal tract size, shape and its dynamics. The distribution of these features obtained using large amount of data is used to represent a given speaker. The distribution is approximated by statistical models such as Gaussian mixture models (GMM) and hidden Markov models (HMM) [1]. These segmental features not only reflect the characteristics of a speaker, but also the characteristics of sound units in a speech signal. To isolate characteristics of a speaker from characteristics of speech sounds is a challenging task.

It is known that the characteristics of speech at the subsegmental level, especially the excitation source, reflects the physiological characteristics of an individual. It is difficult to identify and extract the features at the subsegmental level from speech signal, which contains characteristics of both excitation source and vocal tract system. But it was shown that these source characteristics indeed have speaker-specific information, which is complimentary to the information in the vocal tract system characteristics [2, 3, 4, 5].

The objective of this study is to explore new features specific to a given speaker. These features correspond mostly to subsegmental and segmental features, so that the speaker information can be extracted and represented using a limited amount of data during training, and still lesser amount of data during testing. The main idea is to capture the speaker-specific information from the excitation source and the vocal tract system components of a speech signal separately, as each of these components may contain complementary information characterizing a speaker. In earlier studies [4, 6, 7, 8, 9] auto-association neural network (AANN) models were proposed to capture the speaker-specific information separately from the linear prediction (LP) residual and from the weighted linear prediction cepstral coefficients (wLPCC). In the case of excitation information, the AANN model is used to capture the nonlinear relations among the samples of the LP residual. For this purpose, about 4 ms of the residual samples are presented, with a shift of one sample. The size of the input and the output layers of the AANN models are the same, as the input itself is the desired pattern. In the case of system parameters, a wLPCC vector is presented both as input and the desired output. The wLPCC vectors are derived for each frame of size about 20 ms with a shift of 5 ms. While the complementary nature of the features captured by these two models was demonstrated in speaker recognition experiments [2, 6], the performance of the models individually was not high. It is difficult to understand the information captured by these nonlinear AANN models. Hence it is difficult to find out how the performance can be improved using these models. One reason for poorer performance is probably because all the residual data may not contribute to speaker-specific information. Likewise, it is also likely that extraction of the vocal tract parameters through LP analysis, with a fixed frame size

and a frame shift, may also introduce several spurious wLPCC vectors which may smear the distribution corresponding to the speaker-specific information.

In this paper, we propose the use of pitch synchronous speech data for building AANN models for capturing the speaker-specific information in the excitation source and the vocal tract system. We show that these new models perform significantly better than the models built using pitch asynchronous data as in the traditional block processing using an analysis window of fixed size.

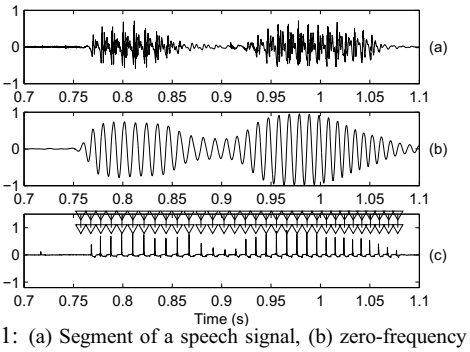

Figure 1: (a) Segment of a speech signal, (b) zero-frequency filtered signal, and (c) differenced EGG signal. Epoch locations marked by arrows.

2. Identification of GCIs for pitch synchronous analysis

The major source of excitation of the vocal tract system in speech production is due to vibration of the vocal folds at the glottis. The instant of significant excitation is due to sharp closure of the vocal folds in each glottal cycle. The glottal closure is almost impulse-like, and the signal energy, and hence the signal to noise ratio (SNR) of speech, is generally high around these instants. Also, some significant speaker-specific characteristics may be present around these instants, as the signal around these regions reflect the vibration characteristics of the glottis of the individual. So by extracting the glottal closure instants (GCIs) from speech signal, it is possible to focus the analysis around these instants to extract speaker-specific information in the excitation and the vocal tract system components of a speech signal.

Recently, a method was proposed to extract GCIs from speech signals using the output of 0 Hz resonator filter. The following are the steps to extract the GCIs [10].

(a) The speech signal $s[n]$ is differenced to remove any slowly varying component introduced by the recording device.

$$x[n] = s[n] - s[n-1] \tag{1}$$

(b) The differenced speech signal $x[n]$ is passed through a cascade of two ideal zero-frequency (digital) resonators. That is

$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n] \tag{2}$$

where $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$. The resulting signal $y_0[n]$ grows approximately as a polynomial function of time.

(c) The average pitch period is computed using the autocorrelation function of 30 ms segments of $x[n]$.

(d) The trend in $y_0[n]$ is removed by subtracting the local mean computed over the average pitch period at each sample. The resulting signal

$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m] \tag{3}$$

is the zero-frequency filtered (ZFF) signal. Here $2N+1$ corresponds to the number of samples in the window used for trend removal. The choice of the window size in not critical as long as it is in range of one to two pitch periods. Fig. 1(b) shows the filtered signal of the speech segment shown in Fig. 1(a). It was shown in [10] that the instants of positive-to-negative zero crossings (PNZCs) correspond to the instants of significant excitation in voiced speech, called *epochs* [10]. The locations of PNZCs of the filtered signal are shown in Fig. 1(c). There is a close agreement between the locations of the strong positive peaks of the differenced EGG (DEGG) signal and the instants of PNZCs derived from the filtered signal. Two pitch periods of the speech signal are chosen for deriving the residual using LP analysis. A 10^{th} order LP analysis is used on the signal sampled at 8 kHz. The system characteristics around each epoch is represented by a 15 dimensional wLPCC vector derived from the 10 LPCs. A 4 ms segment (i.e, 32 samples) of the LP residual is chosen around each epoch to extract the information from the excitation source component.

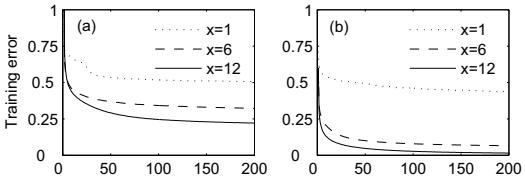

Figure 2: Training error as a function of iteration number, for (a) excitation source models and (b) vocal tract system models. Here x indicates the number of nodes in the compression layer.

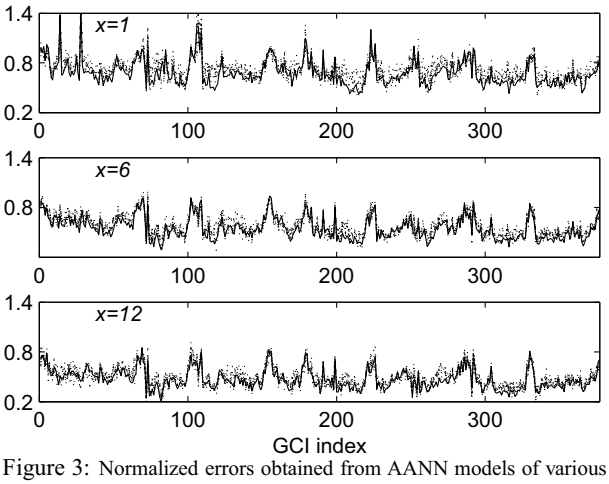

Figure 3: Normalized errors obtained from AANN models of various architectures, using excitation source information. In each plot, solid line ('—') and dotted line ('· · ·') correspond to genuine and imposter curves, respectively.

3. AANN models for capturing excitation source information

A 5-layer AANN model with the structure $32L$ $80N$ xN $80N$ $32L$ is chosen for extracting the speaker-specific information using the 4 ms LP residual around each epoch. Here L refers to linear units, N refers to nonlinear ($tanh()$) output function of units, and x refers to the number of units in the compression layer. The value of x is varied to study its effect on the model's ability to capture the speaker-specific information. The sizes of the input and the output layers are fixed by the number of residual samples (around each epoch) used to train and test the models. The expansion layers provide flexibility for mapping and compression. Typically about 15

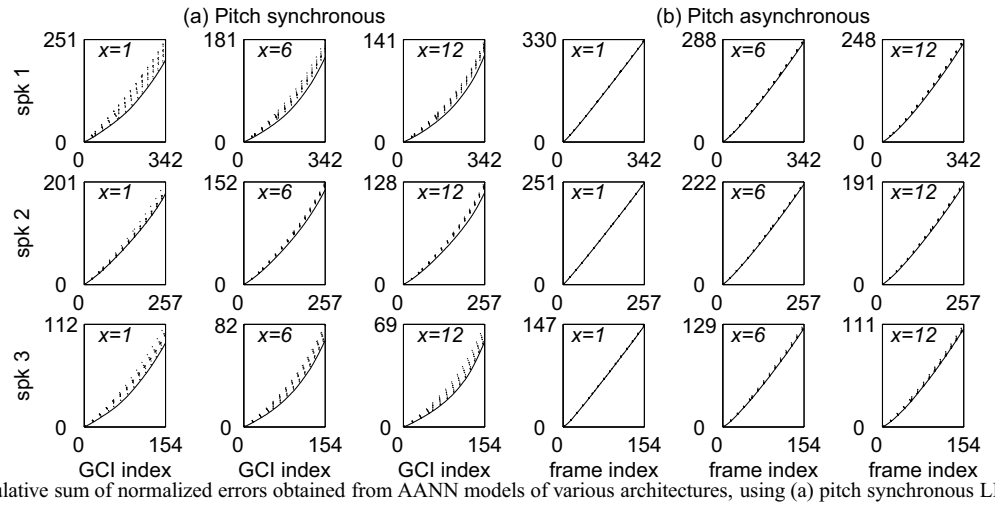

Figure 4: Cumulative sum of normalized errors obtained from AANN models of various architectures, using (a) pitch synchronous LP residual and (b) pitch asynchronous LP residual. In each plot, solid line ('—') and dotted line ('· · · ') correspond to genuine and imposter curves, respectively.

seconds of data is used to train a model for each speaker. Note that only voiced segments are used, as the data is collected only around the GCIs.

The network is trained for 200 iterations, and the training error plots are shown in Fig. 2(a) for different values of the number of units (x) in the compression layer. As can be seen from the plots, since the error is decreasing with number of iterations, the network is able to capture the information in the residual. It is also seen that the decrease in error is more when the number of units in the compression layer are more. But beyond a certain limit on the number, even if the error decreases, the generalizing ability may be poor. Also, the optimal number of units in the compression layer may also be speaker-specific. The effect of the network parameters will be examined in the speaker recognition experiments described in Sec. 5.

4. AANN models for capturing the vocal tract system information

A 5-layer AANN model with the structure $15L\ 40N\ xN\ 40N\ 15L$ is used for extracting the speaker-specific information using 15 dimensional wLPCC vectors. The wLPCC vectors are derived using LP analysis on two pitch period segment around each epoch. The model is expected to capture the distribution of the feature vectors, which is speaker-specific. The training error plots for a speaker for different number of units in the compression layer are shown in Fig. 2(b). The training error plots do indicate that the information in the distribution of the feature vectors is captured. The ability of the model to capture the speaker-specific information can be determined only through speaker recognition experiments, as described in Sec. 5.

5. Speaker recognition experiments

In this section, we discuss the ability of the AANN models described in the previous sections to capture the speaker-specific information, and also the effect of model parameters, especially the number of units in compression layer, on this ability. We have used speech signals from TIMIT database, which consists of 630 speakers and 10 utterances for each speaker. A universal background model (UBM) is built from 100 speakers (50 male and 50 female), using one utterance from each speaker, and the network is trained for 200 iterations. We have used 10 speakers (5 male and 5 female) data for speaker recognition experiments. Eight utterances (approximately 15 seconds of speech data) from each speaker are used to train over the UBM to build the speaker's AANN model, using 200 iterations. For testing, each utterance is presented to a model, and the mean squared error between the output and input, normalized with the magnitude of the input, is computed.

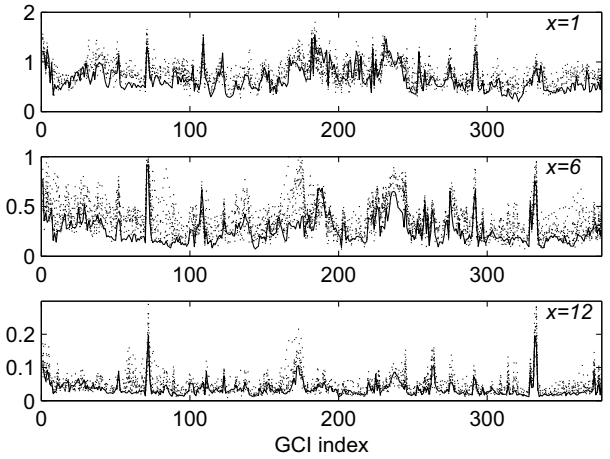

Figure 5: Normalized errors obtained from AANN models of various architectures, using vocal tract system information. In each plot, solid line ('—') and dotted line ('· · · ') correspond to genuine and imposter curves, respectively.

Fig. 3 show the plots of the normalized error obtained from the AANN models of all the 10 speakers at each epoch for a test utterance. The solid ('—') line is the output from the model of the genuine speaker. The test utterance is fed to the models of the other speakers, and the resulting error can be considered as an imposter error. The imposter error curves are shown shown by dotted ('· · · ') lines. The plots correspond to three different values (1, 6 and 12) of the number of units in the middle compression layer. It can be seen that the solid line has the lowest values for most of the frames. The cumulative sum of the error is plotted in Fig. 4(a) to show that the total error is lowest for the genuine speaker. Fig. 4(a) shows the cumulative errors for two other speakers. They show that, for $x = 12$ units in the middle layer, the error for the genuine speaker model is the least for the

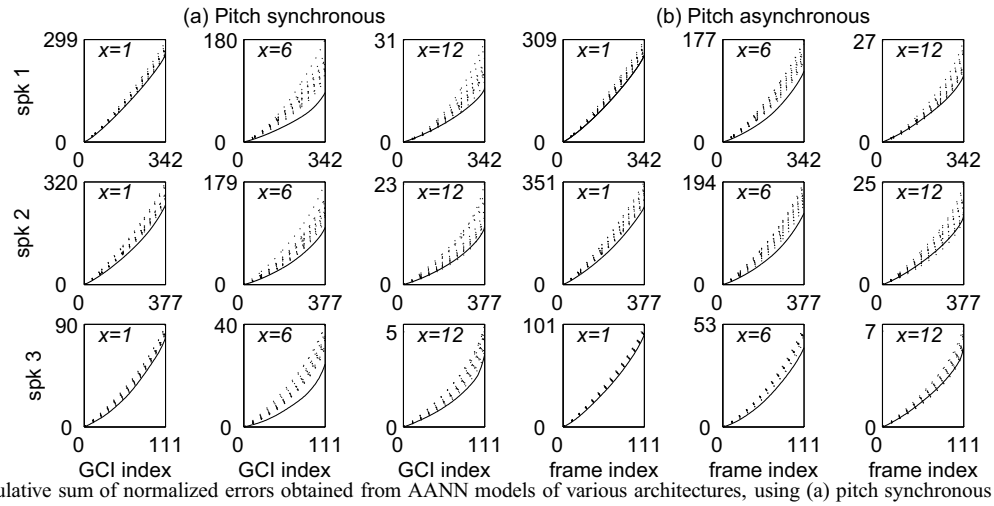

Figure 6: Cumulative sum of normalized errors obtained from AANN models of various architectures, using (a) pitch synchronous wLPCCs and (b) pitch asynchronous wLPCCs. In each plot, solid line ('—') and dotted line ('· · · ') correspond to genuine and imposter curves, respectively.

test utterances. Fig. 4 also shows that for some speakers even a smaller number of units in the middle layer gives equally good performance (spk 1 and spk 3).

Similar observations can be made from Figs. 5 and 6(a) for the error plots for a test utterance tested against all models corresponding to the vocal tract system information using wLPCCs. It is important to note that there is an optimal value for the number of units in the middle compression layer, and this number may be speaker-specific.

Finally, we show that the performance of speaker recognition is inferior when pitch asynchronous data is used for extracting both the excitation and vocal tract system information. Figs. 4(b) and 6(b) show the plots corresponding to cumulative sum of error from AANN models using the LP residual and wLPCC vectors, respectively, derived from a frame size of 20 ms and frame shift of 10 ms. The AANN model for capturing excitation information is trained by LP residual samples obtained from nonoverlapping segments of 4 ms duration. Note that for the purpose of illustration, pitch asynchronous errors (Figs. 4(b) and 6(b)) are resampled to match their lengths with their pitch synchronous counterparts (Figs. 4(a) and 6(a)). For pitch synchronous analysis only voiced frames are used, whereas for pitch asynchronous analysis all the frames of speech are used. Figs. 4(a) (pitch synchronous LP residual) shows good discrimination between genuine and imposter curves compared to Figs. 4(b) (pitch asynchronous LP residual). The difference is less evident for vocal tract information (Figs. 6(a) and (b)).

6. Summary and Conclusions

In this paper, we have demonstrated the significance of using pitch synchronous analysis of speech data and AANN models for extracting speaker-specific information for speaker recognition studies. We have shown that the excitation information is captured using 4 ms LP residual around the GCI, and the vocal tract system information is captured using 15 dimensional wLPCC vectors derived from two pitch period data around each GCI. Speaker recognition experiments were conducted using a subset of speakers from TIMIT data. The results show that features derived from pitch synchronous analysis of speech data give significantly better speaker recognition performance compared to features derived from pitch asynchronous speech data.

In these studies, only the potential of pitch synchronous analysis was demonstrated. Large scale speaker recognition experiments need to be conducted to evaluate the significance of the excitation and the vocal tract system information captured by the AANN models. Also, it is important to explore and develop speaker-specific models by determining suitable AANN models for individual speakers. It is also necessary to explore methods to combine the evidence from excitation source and vocal tract system for speaker recognition.

7. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[2] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Letters*, vol. 13, no. 1, pp. 52–55, Jan. 2006.

[3] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243 – 1261, 2006.

[4] B. Yegnanarayana, K. S. Reddy, and K. Prahallad, "Source and system features for speaker recognition using AANN models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, Utah, USA, May 2001, pp. 491–494.

[5] S. S. Kajarekar, A. G. Adami, and H. Hermansky, "Novel approaches for one and two-speaker detection," in *Proc. European Conf. Speech Proces. and Techn.*, Geneva, Switzerland, Sept. 2003, pp. 2661–2664.

[6] B. Yegnanarayana and K. Prahallad, "AANN: An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, no. 3, pp. 459 – 469, 2002.

[7] S. Joshi, K. Prahallad, and B. Yegnanarayana, "AANN-HMM models for speaker verification and speech recognition," in *Proc. Int. Joint Conf. Neural Networks*, Washington, USA, June 2008.

[8] M. S. Ikbal, H. Misra, and B. Yegnanarayana, "Analysis of autoassociative mapping neural networks," in *Proc. Int. Joint Conf. Neural Networks*, Washington, USA, Dec. 1999.

[9] B. Yegnanarayana, S. R. M. Prasanna, and S. V. Gangashetty, "Autoassociative neural network models for speaker recognition," in *Proc. Workshop on Embedded System, HiPC-2001*, Hyderabad, India, Dec. 2001.

[10] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.