International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-1 Issue-9 March-2013

Speech parameterization based on AM-FM model and its application in speaker identification using AANN

D. Giften Francis Samuel[1], D. Synthiya Vinothini[2]

Abstract

This paper presents the parameterization of speech based on amplitude and frequency modulation (AM-FM) model and its application to speaker identification. Speech parameterization is based on three different bandwidths viz 400Hz, 266mel, 106mel. The feature obtained by this parameterization is termed as PYKFEC which is not directly used as a feature instead its average of each filter is used as the feature and termed as FAP. The speaker identification is done using auto associative neural network and Gaussian mixture model. The AANN/GMM is trained using the SOLO speaking style from CHAINS CORPUS database and a network/model is created for each speaker. The created model is tested using different speaking style like FAST and WHSP of the speaker. The identification rate of FAP is better than PYKFEC, and AANN performs well with these features.

Keywords

Speaker identification, FAP, PYKFEC, AM–FM, amplitude envelope, instantaneous frequency, AANN, GMM.

1. Introduction

Speaker recognition can be classified into speaker identification and verification. Speaker identification is known as closed-set identification since this process determines the best match from the known group of voice for an unknown voice. Speaker verification is known as open-set identification as this process accepts or rejects the identity claim of a speaker, where the unknown voice can also be from an impostor [1]. The base system for speaker recognition system is usually composed of a speech parameterization module, which produces machine readable parameters of the speech samples [2] and a statistical modeling module which creates model representing each speaker.

D. Giften Francis Samuel, Embedded Systems and Design, Raja College of Engineering and Technology, Madurai, India.

D. Synthiya Vinothini, ECE, Karunya University, Coimbatore, India.

This paper introduces a new set of descriptors based on AM-FM representation of speech signal. This new signal characterization is obtained by extending the use of the pyknogram of the signal.

Speech is decomposed using three different band pass setups and feature extracted from each setup is compared and studied with application to auto artificial neural network and Gaussian mixture model. The ability of GMM is to approximate the distribution of the acoustic classes representing broad phonetic events occurring in speech production. The capability of neural network model to discriminate between patterns of different classes is exploited for speaker identification.

2. AM-FM model

AM-FM modulation model represents a single speech resonance r(t) as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) [3],[4].

$$r(t) = \alpha(t) \cos\left[2\pi\left(f_c t + \int_0^t q(\tau)d\tau\right) + \theta\right] \quad (1)$$

where f_c is the "centre value" of the formant frequency, $q(t)$ is the frequency modulating signal, and α(t) is the time varying amplitude. The instantaneous formant frequency signal can be defined as $f(t) = f_c + q(t)$. The speech signal as a whole can be modeled as the sum of *N* such AM-FM signals, one for each formant

$$s(t) = \sum_{k=1}^{N} r_k(t) \quad (2)$$

In order to extract the time varying amplitude (instantaneous amplitude) and instantaneous frequency signals from the speech signal, two steps are followed. Firstly, the speech signal is band passed with a filterbank. Secondly, from the band passed signal, amplitude and frequency modulated signals can be obtained using demodulation schemes like Hilbert transform demodulation.

3. Parameterization of speech signal

Speech parameterization is done by decomposing a input signal with three different filter bank setups. First setup for formant tracking, second setup for

enhancing familiar resonances below 4 kHz and the third setup to approximate the bandwidth scaling of the filters used for extracting MFCCs [5].

A. Filter Bank

Speech signal *s(t)* is passed through a filterbank which has a set of Gabor filters with centre frequencies is uniformly spaced on the hertz scale while the bandwidth remains constant for each setup. A set of bandpass filtered signal $w_i(t)$ is obtained. Gabor filters are optimally compact and smooth in both the time and frequency domain. The bandwidth of Gabor filters used in the first setup is 400Hz and that of the second and third setup is 266 MEL and 106 MEL respectively.

B. Hilbert Transform Demodulation

Demodulation scheme adopted is Hilbert transform demodulation (HTD). HTD has higher computational complexity, but when the first formant frequency is close to the fundamental frequency, it provides smoother estimates for the first formant amplitude and frequency. For each bandpass filtered signal $w_i(t)$, its Hilbert transform $\hat{w}_i(t)$ is computed. Figure 1 shows the block diagram for computing pyknogram.

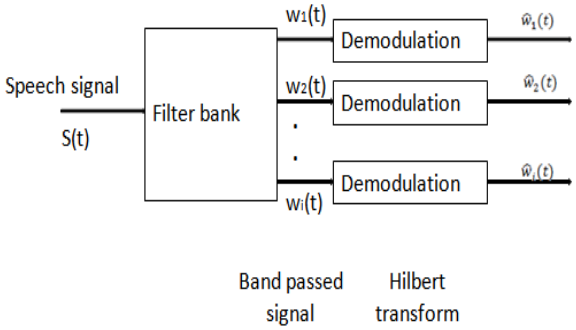

Figure 1: Block diagram for computing pyknogram

C. Pykfec

The instantaneous amplitude $\alpha_i(t)$ and instantaneous frequency $f_i(t)$ for each bandpass signal is computed as

$$\alpha_i(t) = \sqrt{w^2{}_i(t) + \widehat{w^2}_i(t)} \qquad (3)$$

$$f_i(t) = \frac{1}{2\pi}\frac{d}{dt}\left[arc\tan\left(\frac{\hat{w}_i(t)}{w_i(t)}\right)\right] \qquad (4)$$

Instantaneous amplitude and instantaneous frequency are combined together to obtain a mean-amplitude weighted short-time estimate F_i of the instantaneous frequency for each $w_i(t)$.

$$F_i = \frac{\int_{t_0}^{t_0+\tau}[f_i(t)\cdot \alpha_i{}^2(t)]dt}{\int_{t_0}^{t_0+\tau}[\alpha_i{}^2(t)]dt} \qquad (5)$$

This estimate provides a more accurate frequency estimate and it is more robust for low energy and noisy frequency bands [4]. The estimation of this short time estimate leads to extraction of pyknogram of speech signal. Such feature based on the estimates of short-time frequency is referred as pykfec (pyknogram frequency estimate coefficients).

D. Reducing Data Complexity

Pykfec is estimated based on AM-FM model. In this paper, this coefficient is not used as a feature for speaker identification directly due to the complexity of the data dimension. Moreover, difficulties in speech classification stages may arise because speech signals always have different location on the time axis [8]. Therefore, a new algorithm called FAP is proposed in this paper to solve the above mentioned problem.

$X_{FAP}(b_{kj}) = \frac{1}{p}\sum_{i=1+q}^{p+q} b_{ij}$ j= 1,2,….,n. k=1,2..m. (6)

where p=(N/m) represents number of samples from pykfec used in each frame of FAP, q = p*(k-1), b_{ij} represents the pykfec coefficients, N represents the number of time frames in pykfec coefficient, n represents the number of filters used in the filterbank to compute pykfec and m represents the number of frames used in FAP. If k=1, then the two dimensional pykfec coefficient can be reduced to a one-dimensional FAP coefficients. Thus the data complexity is reduced but on the other hand the delay in time axis is ignored. Since the data complexity is reduced it provides a faster and easier method for speaker identification.

4. Classifier

A. Auto Associative Neural Network

AANN model is used for speaker identification using pykfec. AANN models are feed forward neural networks, which performs identity mapping of the input space [6]. From a different perspective, the AANN models can be used to capture the distribution of the input data [7]. Separate AANN models are used to capture the distribution of feature vectors of each speaker. The structure of the AANN model used is 40L 80N 20N 80N 40L, where L denotes linear units and N denotes non-linear units. The activation function of the non-linear unit is the sigmoidal function. The network is trained using error back propagation learning algorithm for 60 epochs. The number of epochs was chosen to obtain

the best performance. The performance of AANN did not improve much, even if the number of epochs was increased.

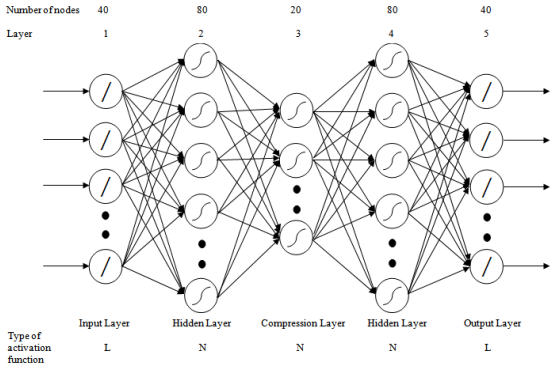

Figure 2: Structure of AANN Model

B. Gaussian Mixture Model

GMM is a simple classifier which is capable of discriminating among different features. During training phase, the algorithm estimates the mixture of Gaussian models that best approximates the distribution of values produced by the speech parameterization module for a given speaker [8]. A statistical model is created for each speaker based on GMM. During testing phase, the feature extracted from an unknown speaker is associated to a speaker model which has maximum posteriori probability, which is obtained by applying Bayes rule and taking logarithm.

5. Database description

Speech samples are extracted from CHAINS corpus. It has recordings of 36 speakers obtained in two different sessions with a time separation of about two months. The first recording session was in a sound proof booth while second one is in a quite office environment. Each speaker provided recordings in six different speaking styles. In this work three different speaking styles are used. NORM: speakers read the text aloud at a comfortable rate; FAST: same text is read at a fast rate; WHSP: same text is whispered [9]. The NORM style that belong to the first recording session is used as a training material while FAST and WHSP from second session is used as a testing material.

6. Results

The speaker identification is done using auto associative neural network and Gaussian Mixture Model. The NORM speaking style signal is used as the training material to train the Gaussian Mixture Model/Auto Associative Neural Network and a model/network is created for each speaker. The trained model/network is then used for testing. The fast and whispered speaking style speech signals are used as testing material. The identification rate using AANN and GMM is calculated for each setup and tabulated. The identification rate is ratio of the number of speakers identified correctly to the total number of speakers.

Table 1: Identification Rate of Setup 1

Identification rate using	Fast Speaking Style		Whispered Speaking Style	
	AANN	GMM	AANN	GMM
FAP	79.3%	69.1%	38.2%	23.5%
PYKFEC	74.2%	63.5%	32.1%	19.7%

Table 2: Identification Rate of Setup 2

Identification rate using	Fast Speaking Style		Whispered Speaking Style	
	AANN	GMM	AANN	GMM
FAP	64.4%	58.5%	30.1%	19.2%
PYKFEC	59.3%	52.7%	24.6%	15.8%

Table 3: Identification Rate of Setup 3

Identification rate using	Fast Speaking Style		Whispered Speaking Style	
	AANN	GMM	AANN	GMM
FAP	88.2%	85.6%	48.2%	44.1%
PYKFEC	83.7%	79.2%	45.4%	39.2%

The identification rate of speakers based on PYKFEC and FAP extracted using first setup, where bandwidth of Gabor filters is 400 Hz is tabulated in Table I. Similarly, the identification rate for speakers based on PYKFEC and FAP extracted using second setup, where bandwidth of Gabor filters is 266 mel is tabulated in Table II and that using third setup, where bandwidth is 106 mel is in Table III.

7. Conclusion

Pyknogram is used to identify the instantaneous frequencies present in speech signal and can be encoded as parameters for speaker identification. The pyknogram reveals the presence of strong resonances

as regions of high density of the estimated short time frequencies. From the experiments, it is clear that the speaker identification rate is better with AANN than GMM for both FAP and PYKFEC. The proposed feature FAP performs well compared to PYKFEC. The features extracted using 3rd setup gives better identification rate than the other two setups.

Acknowledgement

The authors would like to thank M. Ashok Kumar, M.D of Nice fix, currently pursuing his M.Tech in cyber security and cyber law under TIFAC, for his help in publishing our work and also our family members for their support

References

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

[2] F.Bimbot, J.-F.Bonastre, C.Fredouille, G.Gravier, I.Magrin-Chagnolleau, S.Meignier, T.Merlin, J.Ortega-García, D.Petrovska-Delacrétaz, and D.A.Reynolds, "A tutorial on text-independent speaker verification," EURASIPJ. Appl. Signal Process., vol no.1, 2004, pp.430–451.

[3] A.Potamianos and P.Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," J.Acoust.Soc. Amer.,vol.99, 1996, pp.3795–3806.

[4] D.V.Dimitriadis, P.Maragos, and A.Potamianos, "Robust AM-FM features for speech recognition," in IEEE Signal Process. Lett.,vol.12, no.9, Sep.2005, pp.621–624.

[5] Marco Grimaldi, and Fred Cummins, "Speaker Identification Using Instantaneous Frequencies", in IEEE Trans. on Audio, Speech, And Language Processing, Vol.16, No.6, August 2008, pp. 1097-1111.

[6] S. Haykin," Neural Networks: A Comprehensive Foundation", Prentice-Hall International, New Jersey, 1999.

[7] 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.

[8] D.A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," IEEE Trans. Speech Audio Process., vol. 3, no. 1, pp. 72–83, Jan. 1995.

[9] F.Cummins, M.Grimaldi, T.Leonard, and J.Simko, "The chains corpus: Characterizing individual speakers," in Proc.SPECOM'06, St.Petersburg, Russia, 2006, pp.431–435.

D. Giften Francis Samuel was born in Madurai, India, on 24th June 1987. He received the B.Tech degree in Information Technology from Anna University, Chennai, India in 2009. He is currently working as an Assistant Professor in Raja College of Engineering and Technology, Madurai and also pursuing his part time M.E degree in Embedded System Design under Anna University, Chennai, India.

D. Synthiya Vinothini was born in Madurai, India, on 17th March 1988. She received both her B.E degree in Electronics and Communication Engineering and M.Tech degree in Applied Electronics from Karunya University.