Non-Parametric Vector Quantization of Excitation Source Information for Speaker Recognition

Debadatta Pati and S. R. Mahadeva Prasanna
Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati, Guwahati-781039, India
Email: {debadatta, prasanna}@iitg.ernet.in

***Abstract*—The objective of this work is to demonstrate the feasibility of excitation source information obtained by non-parametric Vector Quantization (VQ) for speaker recognition task. Linear Prediction (LP) residual is used as the representation of excitation source information. The LP residual is subjected to non-parametric VQ during training. The codebooks are built for different codebook sizes. The testing of these codebooks using the LP residual of testing speech data indeed demonstrates that a codebook of sufficiently large size uniquely represents the speaker and provides appreciable performance. The speaker recognition system built using conventional Mel Frequency Cepstral Coefficients (MFCCs) representing vocal tract information combines well with the proposed speaker recognition system using excitation source information to provide improved performance. On a set of randomly chosen 30 speakers from the TIMIT database, the proposed system provides 75%, MFCC based system provides 95% and the combined one provides 98.33%.**

***Index Terms*—speaker information, excitation source, vocal tract, VQ**

I. Introduction

Speaker recognition is the task of recognizing speakers using their speech [1]. Depending on whether identity claim is made during testing or not, the goal of speaker recognition can be verification or identification. In verification an identity claim is made along with the testing speech signal. The recognition system will then test speech only against the model of claimed identity and verifies whether the claim is genuine or not. Accordingly, it will either accept or reject the claim. Alternatively, in case of identification only testing speech signal is given to the system and no claim is made. The speaker recognition system will compare the speech with all the models and identifies the most probable speaker of the test speech as the speaker of the model which provides best match. In the present work speaker recognition system operates in the identification mode.

The speaker-specific information present in the speech signal may be attributed to the unique characteristics of vocal tract and excitation source which might have produced that speech. The speaker-specific information related to the vocal tract is manifested mainly as the resonances (formants) and its attributes. There are well developed spectral analysis methods to provide compact representation for the same like Mel Frequency Cepstral Coefficients (MFCCs). The speaker-specific information related to excitation source represents the whole aspect of excitation source which is producing the excitation signal. This includes the change in excitation values from one instant to next, pitch and its variations, instants of significant excitation, glottal wave and its characteristics and so on. However, in the present day systems, excitation source information is attributed mostly to pitch value. Our objective is to look into the excitation source as in independent system and develop methods to represent the signal generated by this system in terms of parameters. These parameters can then be used for speech processing tasks like speaker recognition.

Speech is produced as a result of excitation of time varying vocal tract system using time varying excitation. Mathematically, in time domain, speech signal represents the convolution of excitation signal with vocal tract impulse response. Hence speech needs to be deconvolved to separate and extract excitation source information. One mostly used approach is the inverse filter formulation, where the vocal tract information is estimated first then used in the inverse filter framework to remove the same from the speech signal. After removing the vocal tract information from the speech signal what is left over, more commonly termed as residual will mostly contain the excitation source information [2]. The Linear Prediction (LP) analysis is used to estimate the vocal tract information and remove the same from the speech signal and the resulting signal is termed as LP residual [3].

The LP residual is a signal having less correlation among successive samples and hence difficult to model and represent compactly by the conventional spectral analysis tools. However, it is known both intuitively by human perception studies and also experimental studies that the excitation source part also contains significant speaker information [2] [4], [5]. Hence attempts need to be made to develop methods to model and extract speaker information from the excitation source signal. There are several attempts in the past to model speaker information from the excitation source signal. Atal has demonstrated speaker recognition using pitch contours [6]. Wakita has reported work on using LP residual energy for vowel recognition and also for speaker recognition [7]. Combination of vocal tract based cepstrals with LP residual is shown to give improved performance [8] [9]. Exclusive analysis of speaker specific information from LP residual alone has also been done [4] [5]. Among all these attempts the most recent ones are capturing speaker-specific excitation information using non-linear models like neural networks [2]. The basis for these methods is that since the relations among the samples up to second order have been removed,

the residual may have information only from third or higher order relations among the samples. Such relations may be captured using neural network models. In these studies it is demonstrated experimentally that the neural networks indeed captures speaker information. However, there is always a degree of discomfort to explain what part of excitation is exactly captured in the neural network. In search for the answer, what is proposed in the present work is to directly start from the residual and try to quantize the residual signal itself. This can be achieved by employing the non-parametric approach for Vector Quantization (VQ) [10].

VQ involves quantization of vectors of input speech for compression or pattern recognition [12]. VQ can be either paramteric or non-parametric. In parametric VQ, the speech signal is first represented in terms of feature vectors like MFCC and these MFCC vectors are quantized to prepare the codebook. In non-parametric VQ, the speech signal is directly quantized by considering blocks of raw samples [10]. Of course, parametric VQ provides better compression and less distortion compared to non-parametric VQ. However, when we do not have proper feature extraction techniques, then non-parametric VQ may be a better choice. In the present work, since we are using LP residual which is difficult to parameterize into feature vectors, the non-parametric VQ is used for quantization and representation.

The basis for using non-parametric VQ for representing speaker specific information is that each speaker may have unique excitation sequences manifested in the LP residual and can be captured using VQ. As will be demonstrated through experimental studies later it is indeed the case. The amount of speaker-specific information captured is directly proportional to the codebook size. This is because LP residual is noise like signal and has less correlation among adjacent samples. Hence the unique sequences stored in the VQ may not represent more number of sequences. However, increase in the codebook size will in turn increase computation complexity. Hence it is a trade off between codebook size and amount of speaker-specific information to be captured.

The rest of the paper is organized as follows: Section II describes the development of speaker recognition system based on excitation source information using non-parametric VQ. The development of conventional speaker recognition system using parametric VQ and MFCC feature vectors is described in Section III. The combined speaker recognition system using excitation source information from non-parametric VQ and vocal tract information from parametric VQ is described in Section IV. Summary, conclusions and future scope of the present work are mentioned in Section V.

II. Speaker-Specific Excitation Information using Non-parametric VQ

A. Database

TIMIT database is considered for the present work. It consists of speech data from 630 speakers collected over microphone, sampled at 16 kHz and stored in 16 bits/sample resolution. Out of 630 speakers, 30 speakers are randomly chosen for forming subset for the study. Since most of the speech information is present up to 4 kHz, the speech database of 30 speakers is resampled to 8 kHz and used in the present work. Speech signals of 8 sentences are chosen for each speaker. Out of these, speech signals of first 6 sentences are used for training and speech signals of remaining 2 sentences are used for testing.

B. Excitation Source Information

LP analysis is performed on the speech signal using 10^{th} order prediction [2]. For LP analysis, speech is considered in blocks of 20 ms with shift of 10 ms. For each block of 20 ms, LP Coefficients (LPCs) representing vocal tract information are extracted. The LPCs are used in the inverse filter and speech signal is passed through the inverse filter to obtain the LP residual. Since LPCs characterize vocal tract information, the LP residual mostly contains excitation source information [2].

C. Non-parametric VQ of Excitation Source Information

The LP residual is blocked into frames of 5 ms with shift of 2.5 ms. Longer block sizes of 20 ms will yield code vectors of large size (160 samples at 8 kHz) and hence increased complexity. Alternatively, block sizes of very small sizes like 2.5 ms (20 samples at 8 kHz) may not have enough samples in the sequence for quantization with reliable speaker information due to the noise like nature of LP residual. Hence as trade off between the two, block size of 5 ms is chosen. By considering 5 ms blocks at shift of 2.5 ms, we have about 5500 frames on an average with a standard deviation of about 2000 frames for each speaker.

The 5 ms blocks of LP residual for each speaker are subjected to VQ using LBG algorithm [12]. The codebooks are built for different sizes. For given codebook size, one codebook is built for each speaker. The codebooks are then tested using the LP residuals of the test speech signals considered in blocks of 5 ms with shift of 2.5 ms. The comparison between the codebook and the testing residual segments is made using the Euclidean distance computation. The speaker of the codebook with least average minimum distance is identified as the speaker of the test speech signal.

The speaker recognition system performance for different codebook sizes is given in Table I. The performance is poor for smaller codebook sizes. This is expected since we have large number of frames and are being represented only using few centroid vectors. However, the performance improves significantly as the codebook size is increased. Even though the performance of the largest codebook size is not in par with state of the art systems, as will be demonstrated later, evidence from this system combine well with the conventional speaker recognition system to provide combined improved performance. This exhibits the complementary nature of the speaker-specific information present in the excitation source information.

TABLE I
SPEAKER RECOGNITION PERFORMANCE (%) FOR A SET OF 30 SPEAKERS USING SUBSEGMENTAL FRAMES OF LP RESIDUAL.

Model	Codebook size				
	16	32	64	128	256
Non-parametric VQ	33.33	41.67	51.67	71.67	73.33

III. SPEAKER-SPECIFIC VOCAL TRACT INFORMATION USING PARAMETRIC VQ

A. Database

Same database as used in the earlier study is used.

B. Vocal Tract Information

The Short Term Fourier Transform (STFT) analysis is performed on the speech signal using frame size of 20 ms with shift of 10 ms. For each block of 20 ms, STFT is computed and the magnitude spectrum is further processed by the Mel filter bank to find out the filter bank energies. Then Discrete Cosine Transform (DCT) is taken on the spectral energies to obtain what are called Mel-Frequency Coefficients (MFCCs). For every 20 ms block we get 13 MFCCs excluding c_0.

C. Parametric VQ of Vocal Tract Information

The MFCCs computed using frames of 20 ms with shift of 10 ms are used as parameter vectors for vector quantization. We have about 1700 feature vectors on an average with standard deviation of 250 frames for each speaker. The feature vectors for each speaker are subjected to VQ using LBG algorithm [12]. Since parameter vectors MFCCs are used instead of speech signal samples themselves, this VQ is termed as parametric VQ. The codebooks are built for different sizes. For given codebook size, one codebook is built for each speaker. The codebooks are then tested using the MFCCs of the test speech signal considered in blocks of 20 ms with shift of 10 ms. The comparison between the codebook and the testing segments is made using the Euclidean distance computation. The speaker of the codebook with least average minimum distance is identified as the speaker of the test speech signal.

The speaker recognition system performance for different codebook sizes is given in Table II. The performance is relatively poor for smaller codebook sizes, but it is significantly more compared to the corresponding performance of excitation source information. This is mainly because of the much compact representation of speaker information in the present parametric framework. The performance also improves significantly for large codebook size and is also significantly better compared to the excitation source information based system.

TABLE II
SPEAKER RECOGNITION PERFORMANCE (%) FOR A SET OF FIRST 30 SPEAKERS FOR CONVENTIONAL SYSTEM

Model	Codebook size				
	16	32	64	128	256
Parametric VQ	86.67	90	95	95	95

IV. COMBINING SPEAKER-SPECIFIC EXCITATION AND VOCAL TRACT INFORMATION

The speaker-specific excitation information and vocal tract information represent different aspect of speaker characteristics. As indicated by their name, the speaker characteristics due to the excitation and vocal tract are independent characterizations of the speaker. Hence they may be combined to obtain improved performance [8]. The combination can be done at different levels. In the present case since the LP residual and MFCCs are different measurements, the combination of the two is done at the final level that is, decision level. Accordingly, a non-parametric VQ based speaker recognition system is built using LP residual as described in Section II. A parametric VQ based speaker recognition system is built using MFCCs as described Section III. For the given testing speech signal the LP residual and MFCCs are extracted and testing of the system is done independently. The minimum average Euclidean distance computed in each case are combined using the following relation:

$$E_c = (\frac{P_{ex}}{P_{vt}})(\frac{P_{ex}}{P_{ex}+P_{vt}})E_{ex} + (\frac{P_{vt}}{P_{vt}})(\frac{P_{vt}}{P_{ex}+P_{vt}})E_{vt} \quad (1)$$

where, E_{ex}, E_{vt} and E_c are Euclidean distance of excitation, vocal tract and combined system, respectively and P_{ex} and P_{vt} are the performance of excitation and vocal tract information based speaker recognition systems.

The system with combined minimum average distance is identified as the speaker of the test speech signal. The performance of the system is shown in Tabel III.By comparing Tables I, II and III, it can be observed that the combined system shows an improved performance in each codebook case. This aspect emphasizes the complementary nature of the speaker information present in the vocal tract and excitation parts of speech signal. The speaker recognition study is also repeated for another subset of 30 speakers from the same TIMIT database. The performance for different cases are shown in Tabel IV. In this case also the combined system shows performance which can be attributed to the complementary nature of the speaker information present in the two.

TABLE III
SPEAKER RECOGNITION PERFORMANCE (%) FOR A SET OF FIRST 30 SPEAKERS FOR COMBINED NON-PARAMTERIC AND PARAMETRIC VQ SYSTEM.

Model	Codebook size				
	16	32	64	128	256
Combination	86.67	95	98.33	98.33	98.33

TABLE IV
SPEAKER RECOGNITION PERFORMANCE (%) FOR ANATHOR SET OF 30 SPEAKERS

Model	Codebook size				
	16	32	64	128	256
Non-Parametric	40	48.33	56.67	66.67	73.33
Parametric	88.33	90	93.33	93.67	93.33
Combination	90	88.33	95	95	95

V. Summary and Conclusions

This work aimed at exploring the usefulness of non-parametric VQ for modeling speaker-specific excitation information. A speaker recognition system using the same was built and showed appreciable performance for large codebook sizes. A speaker recognition system was also built using vocal tract information by employing parametric VQ. Finally the two systems were combined and combination showed improved performance. This demonstrated the complementary nature of speaker-specific excitation information modeled by non-parametric VQ.

In this work it is demonstrated that non-parametric VQ can indeed be used for modeling speaker-specific excitation information. The present work needs to be extended for larger database size. The non-parametric VQ concept may also extended for extracting speaker-specific excitation information at other levels.

References

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

[2] S. R. M. Prasanna, C. S. Gupta and B. Yegnanarayana, "Extraction of speaker information from exciation source", Speech Communication, vol.48, pp.1243-1261, Oct., 2006.

[3] Makhoul J., "Linear Prediction: a tutorial review", Proc.IEEE, pp.561-580, Oct., 1975.

[4] Liu J. H. L. and Palm G., "On the use of features from prediction residual signal in speaker recognition", Proc. Int. Conf. European Conf. Speech Tecognology (EUROSPEECH), 1997.

[5] Thevenaz P. and Hugli H., "Usefulness of LPC residue in text-independent speaker verification", Speech Communication, vol.17, pp.145-157, 1995.

[6] B.S.Atal, "Automatic Speaker Recognition Based on Pitch Contours",J.Acoust.Soc.am.,Vol.52,pp.1687-1697,1972.

[7] Wakita H., "Residual energy of linear prediction to vowel and speaker recgontiion", IEEE Trans. Acoustics, Speech, Signal Processing, vol.24, pp.270-271, 1976.

[8] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah and C. S. 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, July 2005.

[9] B. Yegnanarayana, Reddy K. S. and Kishore S. P., "Source and system features for speaker recognition using AANN models", Proc. Int. Conf. Acoustics, Speech, Signal Processing, 2001.

[10] V. Cuperman and A. Gersho, "Vector Predictive Coding of Speech at 16Kbits/s", IEEE Trans. Communication, vol.COM-33, no.7, pp.685-696, July 1985.

[11] Mermelstein P. and Davis S. B., "Comparison of parametric representations for monosyllabic word recognition in continuosly spoken sentences", IEEE Trans. Acoustics, Speech, Signal Processing, vol.28, pp.357-366, 1980.

[12] Y. Linde, A. Buzo and R. Gray, "An Algorithm for vector Quantization design", IEEE Trans. Communication, vol.28, pp.84-95,1980.