IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 1, July 2010
ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

Analysis of FFSR, VFSR, MFSR Techniques for Feature Extraction in Speaker Recognition: A Review

Prof. Rupali Pawar, Miss. Hemangi Kulkarni

[1] Department of Computer Engineering, Viswvakarma Institute of Technology, Pune University Pune, Maharashtra, 411037, India

[2] Department of Computer Engineering, R.H. Raisoni College of Engineering and Management, Pune University Pune, Maharashtra, 411037, India

Abstract

In this paper we provide the brief overview of FFSR, VFSR and MFSR technique for speech analysis in speaker recognition system. Following this overview we will discuss some of the strength and weakness of current frame size and frame rate technique. A Fixed Frame Size and Rate (FFSR) in most of state-of–the-art speech recognition system can face some problems such as accidentally meeting noisy frames, also assign the same importance to each frames. As an attempt to avoid these problems Variable Frame Size and Rate (VFSR) technique selects optimal frame size and frame rate depending on speaking rate to capture sudden changes in spectral information along with time. But it increases the burden of identifying the spectral changes in the speech. To overcome this drawback Multiple Frame Size and Rate (MFSR) is used which provide different feature vectors for same data which increases the performance of speech recognition. The motivation of all these analysis is to increase the speech recognition performance at the cost of reducing the time and space complexity.
Index Terms: *Fixed Frame Size and Rate (FFSR), Variable Frame size and Rate (VFSR), Multiple Frame Size and Rate (MFSR).*

1. Introduction

Speaker recognition is the process of automatically recognizing who is speaking by using speaker specific information included in speech waves (Dod- dington, 1985; Furui, 1986, 1989, 1991a,b, 1994; O'Shaugnessy, 1986; Rosenberg and oong, 1991). This technique can be used to verify the identified claimed speaker's voice. Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. In most of the applications voice is used to confirm the identity claim of a speaker .

Speaker recognition system may be viewed as working in four stages namely Analysis, Features Extraction, Modeling and Testing as shown in Fig.1.

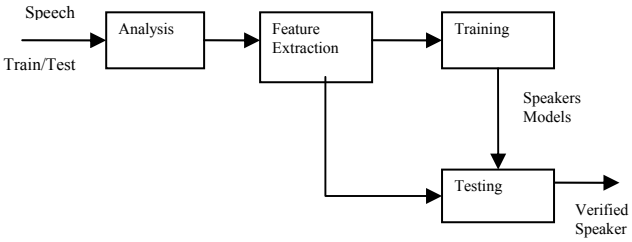

Fig. 1: Stages in the Development of Speaker Recognition System.

Speech data contains different types of information that conveys speaker identity. These include speaker specific information due to the vocal tract, excitation source and behavioral traits. The speech signal is produced from the vocal tract system. The physical structure and dimension of vocal tract as well as the excitation source are unique for each speaker. This uniqueness is embedded into the speech signal during speech production and can be used for speaker recognition. To obtain the good representation of these speaker characteristics, speech data needs to be analyzed. The speech analysis stage deals with the selection of suitable frame size and frame shift for segmenting the speech signal for further analysis and feature extraction.

The speech analysis is done using one of the following techniques: Segmental analysis, Sub-Segmental Analysis and Supra-Segmental analysis [11] as shown in Fig. 2.

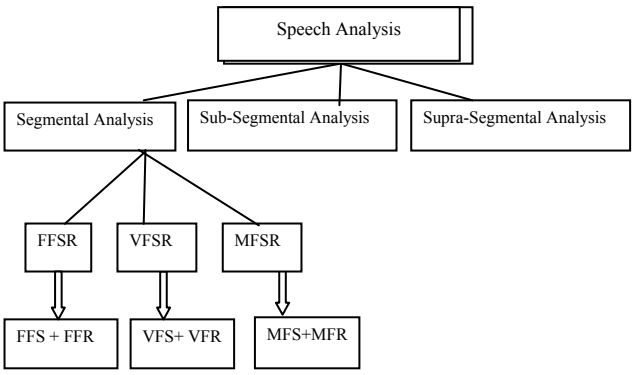

Fig. 2: Different Types of Speech Analysis Depending on Frame Size.

The most feature extraction techniques extract useful features from short-time spectra. To get the short-time spectra, an utterance is segmented into a number of frames. Two parameters, including frame size and shift period (rate), are involved in segmenting the utterance. Window size is the length of each frame. Shift period is the length of time between the starts of two consecutive frames.

The two parameters in segmenting the utterance, frame size and shift period (rate) affect the time requirement in feature extraction. Let W is the window size for an utterance with the L length of signal in seconds, and S is the shift period. Then their relationship can be expressed as

Number of feature vectors = 1+ (L-W)/S (1)

W must be kept larger than S, otherwise some segments of utterance are not analyzed. When W equals S, there is no overlapping area between each frame. Moreover, when S increases, the time taken in feature extraction also becomes shorter. Not only the time taken in feature extraction, size of the shift period also affects the time taken in training a template model because when the shift period decreases, the number of extracted feature vectors increases. Also changing S and W affects the verification accuracy. Window size larger than the critical point not only increases the computational effort but also reduces the verification accuracy. On the contrary, using a window size less than the critical point results in tradeoffs between accuracy and computation time [1].

The speech signals are non- stationary and exhibit quasi-stationary behavior at shorter durations after segmentation in speech analysis. The conventional speech recognition systems use features that are extracted with fixed frame size and frame rate (FFSR) also known as single frame size and frame rate (SFSR). In most of the cases, such feature extraction works well. But this may face problem when the test speaker's pitch frequency and/or speaking rate is very different from that of the speaker's data used during training. Another problem with conventional single frame size feature extraction is that it may not be able to capture the sudden changes in the spectral information along time. To overcome these limitation of FFSR, a new Variable frame size and rate (VFSR) techniques described in [4], [5], [6] allow us to adjust the frame rate according to rate of change of spectral information along time. But it requires additional computation time. To reduce the computation complexity and time in capturing the sudden changes in the spectral information, a new technique Multiple Frame Size and Rate (MFSR) [9] is used. In MFSR, same speech data is analyzed at different frame sizes and rates, the set of speech samples involved in the analysis are different at each frame size and rate and hence the feature vectors representing the vocal tract information may be different .Thus the MFSR analysis technique generates more number of feature vectors and hence may result in better speaker modeling and testing [9].

2. Effect of Frame Shift in Frame Selection

In typical frame-based speech signal processing, the frame shift is fixed to about half of the frame length for practical consideration, Selecting feature frames for speaker recognition in which the frame shift is fixed to around half of the frame length may not be the best choice, because the characteristics of the speech signal may rapidly change, especially at phonetic boundaries. However, such an approach does not reflect the dynamic characteristics of spectral features that are caused by co-articulation within a single analysis frame.

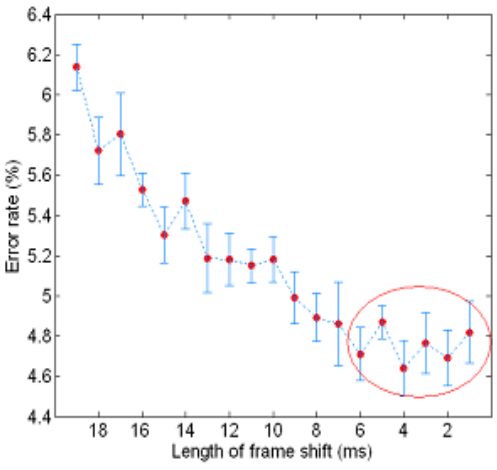

Fig. 3: Shows that the error rate generally decreases as the frame shift decreases.

The conventional method that uses a fixed frame shift is considered inefficient because it does not take into consideration the distribution of speech samples among different phoneme classes. By varying the frame shift based on the given phonetic information, the performance of speaker recognition systems can be improved. Increasing the number of feature frames by decreasing the frame shift results in performance improvement. The error rate generally decreases as the frame shift decreases.

Compared to the conventional method of half frame shift, the improvement is statistically meaningful. However there is a limit to improve the recognition performance by only increasing the amount of redundant feature frames in utterances derived. Furthermore, mutual information directly correlates with recognition performance, a new feature frame selection method based on the normalized minimum-redundancy and maximum-relevancy (NmRMR) [3] criterion, which minimizes redundant information between selected feature frames but maximizes mutual information between speaker models and test feature frames.

3. Effect of Frame Size and Frame Rate in Frame Selection

The applied features extraction and the modeling approaches may be too sensitive to intrinsic speech variability's, like: speaker, gender, age, dialect, accent, health condition, speaking rate, prosody, emotional state, spontaneity, speaking effort, articulation effort. The performance of the speech recognition system depends on the local conditions within an utterance for each speaker. It is important to capture this local variation within an utterance. In order to capture this information FFSR, VFSR and MFSR techniques are used in feature extraction.

3.1. Fixed Frame Size and Rate Analysis on Speech

For speech signals, some regions are relatively constant (stationary speech) while others may change rapidly in time (dynamic speech). In most of the speech processing systems, speech signals are first windowed into frames. Frames are typically 20-30ms in duration and the frame shift is 10ms. The justification for such segmentation is that the speech signals are non- stationary and exhibit quasi-stationary behavior at shorter durations. Fixed frame size with keeping the frame rate fixed for segmentation can be consider as fixed frame size and rate(FFSR)

3.1.1. FFS

Most speech recognition systems use a fixed frame size, to decompose speech into a series of frames, basically due to its convenience. Fixed-frame windowing results in better recognition rates with clean signals. While when noise is introduced to the system, performance degrades [10]. Speech frames are often assigned the same importance in pattern classification. However, in the case of continuous speech recognition, observations at the beginning and end of a phoneme are highly influenced by contextual information. The frames at boundaries may carry more speaker-related information, and are often vague for the speech recognition task. At the same time observations in the steady zone add redundant information in the decision process. A ML-based frame selection technique[12], which selects reliable frames from a tiny frame shift, to classify phonemes by assuming the boundaries of phonemes are known priory.

The use of fixed frame representations has several advantages. It provides a stable stream of information about the speech signal without being affected by the underlying signal content. Fixed frame representation provides systematically information of the spectral content in the transient signal. Temporal resolution of fixed frames is also good if the window step size is sufficiently small, which means that the quantized label sequences can describe short-term details in the signal.

3.1.2. FFR

Commonly agreed that a fixed frame rate, typically 10ms, is not consistent with human perception, the fixed frame rate is still used in most of state-of-the-art speech recognizers because of its simplicity and convenience. A fixed frame rate adopted in most state-of-the-art speech recognition systems can face some problems, such as accidentally meeting noisy frames, assigning the same importance to each frame, and pitch a synchronous representation. As an attempt to avoid those problems, ML approach [12] selects reliable frames from a fine resolution along the time axis. But some limitations appear with this arbitrary and fixed frame rate. For instance, a noisy frame may dominate the recognition process; the same importance is assigned to each extracted frame, which has been shown in consistent with human perception. Besides, pitch asynchronous representation caused by the fixed frame rate leads to pitch mismatch due to the presence of pitch related harmonics in the power spectrum. Because of those limitations, two methods that are less sensitive to frame position, are proposed to select reliable frames along the temporal axis, which are both based on an ML criterion[12]. The first method, called multiple selections

selects frames from a tiny frame shift after which the average frame rate matches a pre-defined value. The procedure can be incorporated in to a modified Viterbi decoding algorithm. The other method, named single selection, selects only one frame for each state of a Hidden Markov Model (HMM). In the single frame selection, it allows to select a single frame for each state dynamically. Single selection is capable of selecting more suitable frames from the pool of frames with a finer time resolution, and hence achieves superior performance. It is quite possible that frames selected by the multiple selections are still redundant and useless, or even weakening because of the constraint of the intended average frame rate; some of them can be discarded further. Multiple selections to continuous speech recognition are much easier than the single selection as its implementation is very similar to classical time-synchronous Viterbi. The single selection needs the hypothesis of phoneme boundary [12].

3.1.3. FFSR

Representing a segment of speech, with effective resolution depends on both frequency and time position within the segment. Tradeoffs between time and frequency resolution can be made by varying the frame size and frame spacing. Under a constraint of fixed frame rate, a practical approach is to code the dynamic speech with higher time resolution by analyzing it with shorter frame and code stationary speech in higher frequency resolution by analyzing it with longer frame. An adaptive frame length selection approach (AFL) [14] can adapt to the frequency/time resolutions of the transform depending upon the spectral and temporal characteristics of the signal being processed. During the feature extraction process, each speech frame is examined at different time scales. If a transient is detected in the second half of a speech frame, then this frame will be analyzed with short frame(a normal frame is divided into two short frame of same length in the experiments).Thus, more accurate speech coding can be achieved by using the adaptive frame length method. It is expected that this accurate coding of speech could result in better recognition performance.

In most of the cases, using FFSR, feature extraction works well. But, this may face problem when the test speaker's pitch frequency and/or speaking rate is very different from that of the speaker's data used during training. Another problem with conventional fixed frame size feature extraction (for example, with a frame size of 20 ms) is that it may not be able to capture the sudden changes in the spectral information along time [8].

3.2. Variable Frame Size and Rate Analysis on Speech

Changes in spectral characteristics are important cues for discriminating and identifying speech sounds. These changes can occur over very short time intervals. Computing frames every 10 ms, as commonly done in recognition systems, is not sufficient to capture such dynamic changes. One solution to this problem is increasing the frame rate, but this would unnecessarily increase the computational load of ASR systems and is not needed for steady segments. Instead of this a variable frame size and rate method in which the frame size and rate varies as a function of the spectral characteristics of the signal is used. A Variable Frame size and Rate (VFSR) technique results in an increased number of frames for rapidly-changing segments with relatively high energy and less frames for steady state segments[4],[5].

3.2.1. VFS

The use of VFS to handle phenomenon such as speaking rate has been explored in the context of speech coding as well as ASR. For better modeling of transition segment, the size of analysis frame depends on the data that frame contains and it varies within an utterance. This means size of analysis frame should be reduced near transition region. Two VFS analysis methods wide -band analysis and narrow-band analysis are used [4].

i. Wide-band analysis method is used for short duration (typically 3ms) analysis frame to achieve high time resolution and for larger frame size (typically 25ms) to achieve high frequency resolution.
ii. Narrow –band analysis is used to represent resonant characteristics of steady regions of continuant sound such as vowel.

The wide-band analysis does not lose much fine spectral features because the time average of varying fine features appears as gross feature, also high time resolution results in better representation of temporal trajectory of system. This will increase recognition accuracy [4].VFS analysis method has two steps:

i. Detection of dynamic region of speech signal.
ii. Change of frame size to suit the local condition.

3.2.2. VFR

Using FFSR, it is clear that computing frames at fixed rate(typically 10 ms) is not adequate for representing rapidly changing segments although it is sufficient for representing relatively steady and long ones. The solution to this problem is increasing the frame rate, but this would unnecessarily increase the computational load of ASR

systems and is not needed for steady segments. Instead, a variable frame rate method in which the frame rate varies as a function of the spectra1 characteristics of the signal is used [5].

Several strategies which employ frame rates that vary within an utterance for automatic speech recognition have been reported in the literature. They recognize the fact that some frames in the regions can be dropped; the detection of such steady segments is based on Euclidean Distance between current and previous frames [Bridle and Brown, 1982]. If the distance of frame to its neighbor is less than threshold, the current frame discarded [4]. A new variable frame rate analysis based upon the first-order difference of the log energy for each frame (ΔE) [6]. Compared with previous variable frame rate methods, this delta energy approach is simpler and achieves similar recognition accuracy improvements but at reduced complexity. This ΔE VFR analysis relies upon criterion to determine at what point a new feature should be extracted. Intuitively, new features should be extracted only after sufficient changes have occurred within the speech signal to warrant their extraction. Previously[4],[5] some kind of feature-based measure have been compared against a pre-determined threshold in order to determine whether a particular feature should be retained or not, however in principle any criterion that yields similar discriminating power can be used. The criterion employed in [6] is to retain the current frame if the change in energy ΔE is greater than a fixed threshold T, and discard it if ΔE < T,

$$\Delta E = E_m - E_{m-1}, \text{ and } E_m = \log\left[\sum_{n=0}^{N} x_m^2(n)\right] \quad (2)$$

Where m is the frame number, N is the frame length and xm (n) is the nth sample of speech in the mth frame. Two methods [6] for calculating ΔE are possible for a VFR approach:

i) ΔE is based on the log energy difference between consecutive frames with a fixed spacing.
ii) ΔE is based on the log energy difference between the last retained frame and the current frame.

ΔE VFR analysis improve recognition of speech in the presence of babble, car or exhibition noise, particularly where the test data are noisier and more diverse than the training data. This approach is faster than existing methods for VFR speech analysis as it does not require candidate features to be pre-computed [6].

3.2.3. VFSR

VFSR provide the advantages over FFSR like reduction in the number of processed frames without degradation in the performance, better acoustic signal modeling in regions with fast spectral changes and increased immunity to performance loss in task suffering from additive noise. But calculating the rate of change of spectral information along time requires additional computation time.

3.3. Multiple Frame Size and Rate Analysis on Speech

It is observed that the feature vectors representing the vocal tract information extracted from the same speech signal by multi-resolution analysis are considerably different. Further, the speaking rates as well as pitch are different for different speakers and also for the same speaker depending on the contextual information during speech production. In MSFR the same speech is analyzed using different frame size and rate and hence it is termed as multiple frame size and rate (MFSR) analysis.

3.3.1. MFS

In MFS analysis, speech data is analyzed using different frame sizes but with constant frame shift. It is effectively a multi-resolution analysis technique. The magnitude spectra and the resultant feature vectors extracted from the speech signal with different frame sizes are considerably different due to different frequency resolutions [2]. This is because the information presenting the spectral domain is due to the convolution of true spectrum of speech and spectral domain window. Further, the speech samples in each frame size are slightly different .Both these factors may lead to varied levels of manifestation of speaker information in different feature vectors. Thus by varying frame size, we can vary the spectral information manifested and hence the feature vectors with different speaker-specific information. MFS analysis results in an increased number of feature vectors for the same speech data [9].

3.3.2. MFR

In MFR analysis, same speech data is analyzed using different frame shifts (rates) with constant frame size. The speaking rates as well as pitch are different for each speaker. This is because speaking rate is a behavioral aspect of speaker information, which depends on how the speaker is habituated to produce speech. The pitch is an attribute of the excitation source related with the speaker, and these two information vary for the same speaker depending on the contextual information during speech

production. Since speaking rate and pitch are different, rate of change of spectral information will be different. Therefore by analyzing the same speech data at different frame rates, set of speech samples involved in the analysis of speech are different at each rate and hence feature vectors representing vocal tract information may be different. Thus MFR analysis is effectively a multi-shifting technique [8]. Accordingly, spectral resolution will remain the same, but there will be new set of speech samples for each shift [9].

3.3.3. MFSR

The main advantage of MFSR over VFSR is reducing the burden of identifying the spectral changes in speech. In MFSR analysis, same speech data is analyzed using both MFS and MFR analysis techniques, the speech samples involved in the analysis are different at each frame size and rate and hence the feature vectors representing the vocal tract information may be different. Thus the MFSR analysis technique generates more number of feature vectors and result in better speaker modeling and testing.

4. Conclusion

It is known that performance of speaker recognition depends on the speaking rate of speaker and it vary for different speaker. If the test speaker's speaking rate is different from that of trained speaker's speaking rate then it affects the performance of speaker recognition with respect to time, space and computational complexity. This paper describes the three techniques for selection of frame size and frame rate to reduce the time, space and computational complexity. First, FFSR keeping frame size and frame rate constant throughout the experiment gives better performance for steady state speech, but it may not be capable to capture sudden changes in the spectral information along time. Instead of this, Second, VFSR where frame size and frame rate vary with respect to change of spectral information provide better performance when there is sudden change in speech. But it increases the time and computational complexity for finding such changes in speech spectrum. To overcome this problem, Third , MFSR where multiple frame shift and multiple frame rate technique is used which results increased number of frames and reduces the computational complexity and time without calculating the rate of changes of spectral information along time.

References

[1] C.C. Leung and Y.S. Moon, 'Effect of Window Size and Shift Period in Mel-Warped Cepstral Feature Extraction on GMM-Based Speaker Verification', J. Kittler and M.S. Nixon (Eds.): AVBPA 2003, LNCS 2688, pp. 438-445

[2] Lawrence Rabiner, B H Juang, Biing Hwang Juang, 'Fundamentals of Speech Recognition',(Prentice Hall, Singapore) , ISBN: 0130151572

[3] Chi-Sang Jung, Moo Young Kim, and Hong-Goo Kang, 'Selecting Feature Frames for Automatic Speaker Recognition Using Mutual Information', Audio, Speech, and Language Processing, IEEE Transactions on Volume: PP , Issue: 99 IEEE 2009 , Page(s): 1 – 1.

[4] Samudravijaya,K.,(2004),'Variable frame size analysis for speech recognition', Proceedings of Int. Conf. Natural Language Processing, New Delhi, India, pp. 237-244

[5] Qifeng Zhu; Alwan, A, 'On the use of variable frame rate analysis in speech recognition,' in Proc. IEEE ICASSP, 2000.86209 , Page(s): 1783 - 1786 vol.3.

[6] Le Cerf, P., and Van Compernolle, D., 'A new variable frame rate analysis method for speech recognition,' IEEE Sig. Proc.Letters, vol.1, no.12, pp. 185–187, December 1994

[7] Ponting, K.M. and Peeling, S.M.'The use of variable frame rate analysis in speech recognition.' Computer Speech and Language Comput. Speech Lang. (UK), vol.S, (no.2), April 1991. p.169-79.

[8] Saradag L., Nagarajan T.,Murthyh A.,' Multiple frame size and multiple frame rate feature extraction for speech recognition' .Proc.Int.Conf. Signal Process. Communication, Bangalore, India, October2004

[9] H. S. JayannaS.R.MahadevaPrasanna, 'Multiple frame size and rate analysis for Speaker recognition under limited data condition', IETSignalProcess., 2009, Vol.3, Iss.3, pp.189–204189

[10] Okko Räsänen , Joris Driesen ,'A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition', Nordic Conference of Computational Linguistics NODALIDA 2009, NEALT Proceedings Series, Vol. 4 (2009), 255-262.

[11] H. S. Jayanna and S. R. Mahadeva Prasanna, 'Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition', IETE Technical Review, Vol 26, Issue 3, May-June2009

[12] Tingyao Wu,Van Compemolle D,Duchateau J,Van Hamme H; 'Maximum Likelihood Based Temporal Frame Selection',Acoustic, Speech and Signal Prosessing, ICASSP 2006 Proceeding,2006 IEEE Vol 1,Page(s): I-I

[13] T.Y. Wu, D.Van Compernolle, J. Duchateau, H.Van Hamme, 'Single frame selection for Phoneme Classification'. Proceedings of. International Confe-rence on Spoken Language Processing, Pittsburgh, USA, September 2006, 641-644.

[14] SamKwong,QianhuaHe, 'The Use of Adaptive Frame for Speech Recognition', EURASIP Journal on Applied Signal Processing2001:2,82–88.

Prof. Rupali V Pawar
Working as Assistant Professor in Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, India. She has completed her ME in Electronic Engineering , from Government College of Engineering , University of Pune, India, in year 2006 . Research Scholar for PHD in Speech Processing. Published IEEE paper on speech recoginition. Area of her interest is: DSP, Speech processing

Miss. Hemangi S Kulkarni
Working as Lecturer in R.H. Raisoni College of Engineering & Management, Pune. Research Scholar of ME (CSE-IT) from Vishwakarma Institute of Technology, Pune, Maharashtra, India. She has completed her BE in Computer Engineering , from Pune University , India, in year 2003 and ME in Computer Science and Engineering from VIT,Pune in year 2010 .