MULTIPLE FRAME SIZE AND MULTIPLE FRAME RATE FEATURE EXTRACTION FOR SPEECH RECOGNITION

G. L. Sarada, T. Nagarajan, and Hema A. Murthy

Department of Computer Science and Engineering
Indian Institute of Technology, Madras
sarada,raju,hema@lantana.cs.iitm.ernet.in

ABSTRACT

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 and dynamic changes within an utterance using CDHMMs, we propose a novel approach where the features of each utterance are extracted using multiple frame sizes and multiple frame rates where models are trained with these features using CDHMM. The performance of the recognition system using multiple frame size (MFS) and multiple frame rate (MFR) feature extraction is compared with that of the single frame size, where the window size and frame rate are fixed. Using this approach, for a gender dependent speech recognition system, there is an observable improvement in the performance of 4% over the recognition system using single frame size feature extraction.

1. INTRODUCTION

In most of the speech processing systems, speech signals are first windowed into frames. Frames are typically 20-30ms in duration and the frame step size is 10ms. The justification for such a segmentation is that the speech signals are non-stationary and exhibit quasi-stationary behavior at shorter durations. The conventional speech recognition systems use features that are extracted with single frame size and frame rate. 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 (for example, with a frame size of 20 ms) is, it may not be able to capture the sudden changes in the spectral information along time. Variable frame rate techniques described in [1],[2] and [3], allow us to adjust the frame rate according to the value of some specific metric. In such techniques, the frame rate is proportional to the rate of change of spectral information along time. The major claims in using such techniques are the following:

- Reduction in the number of processed frames without degradation in the performance.
- Better acoustic signal modeling in regions with fast spectral changes.
- Increased immunity to performance loss in task suffering from additive noise

Firstly, with the use of the modern computing facilities, increased number of frames to be processed may not be a serious issue. Secondly, calculating the rate of change of spectral information along time, requires additional computation time. Instead, if a fixed, and increased frame rate is used for feature extraction may give a better acoustic modeling.

Here the motivation is to reduce the computation complexity and time in capturing the sudden changes in the spectral information, without calculating the rate of change of spectral information along time.

In our lab, for language identification task [4] and automatic transcription task [5], an unsupervised incremental approach is proposed. In [4] and [5], multiple frame size feature extraction is used for initializing models and compared with single frame size feature extraction and claimed that multiple frame size (MFS) feature extraction can handle variations in features. But, in these experiments, the data used for testing is same as train data.

In this paper, an attempt is made to check the effect of multiple frame size feature extraction, on a different test data. When the number of training examples is less, it is noted that MFS feature extraction outperforms SFS feature extraction. To analyze the performance of the effect of MFS on different amount of training data, another experiment is also carried out. Further, with an assumption that the variation in the speaking rate of the test speaker can be captured by multiple frame rate (MFR), another experiment is carried out with MFR feature extraction. In all these experiments, the main focus is on comparing the performance of the proposed feature extraction techniques with single frame size feature extraction. Since the focus is only on relative performance, specific care is not given in tuning the model parameters. Only acoustic likelihood based raw recognition results are analyzed in all the experiments.

The organization of this paper is as follows. In Section 2, the MFS feature extraction and its effect on CV unit recognition is described. In Section 3, MFS combined with MFR feature extraction and their effects are discussed in detail.

2. MULTIPLE FRAME SIZE FEATURE EXTRACTION

Multiple frame size based feature extraction technique is used in [4] for language identification task, where the models are initialized with a single training example. In that work [4], it was argued that Multiple Frame Size (MFS) feature extraction ensures a reasonable variance for each Gaussian mixture in the models. To show the effect of MFS feature extraction, an experiment is conducted as explained below. For this experiment, 200 syllable segments are considered for training and the same set is used for testing also. In the case of MFS feature extraction, the features are extracted with 5 different frame sizes (12,14,16,18,and 20 ms). For Single Frame Size (SFS) feature extraction, the features are extracted with single window size (20 ms). Hidden Markov Models for the 200 syllable segments are initialized in both SFS and MFS separately. During testing, from the same syllable segments, the features are extracted with 9 different frame sizes (13,15,17,19,20,21,23,25,and 27 ms) and tested against both (SFS and MFS) set of syllable models. The performance is given in the Figure 1. This shows that, models generated with MFS features can handle variations (see the performance for 21,23,25, and 27 ms features).

In the above described experiment, for both training and testing same data is considered. To see the effect of MFS feature extraction on a different speaker data during testing, an experiment is carried out as explained below.

The Indian language television news database [6] is used for the following experiments. The Tamil language news bulletins, consisting of five female speakers, are considered for training. Each speakers data is segmented into syllable-like units and the similar syllable segments are grouped together manually. The syllables which contain CV units alone are taken for this experiment. Presently, for analysis 40 different CV units are considered, for which enough training examples are available. All the similar syllables are grouped together, and features (13 dimensional MFCC + 13 delta + 13 acceleration) are extracted using different frame sizes (12, 14, 16, 18 and 20ms) and models (3 states, 1 mixture/state) are generated. These models are tested against a different female speaker's speech data and the results are shown in Figure 2. Here the models are generated using female data alone.

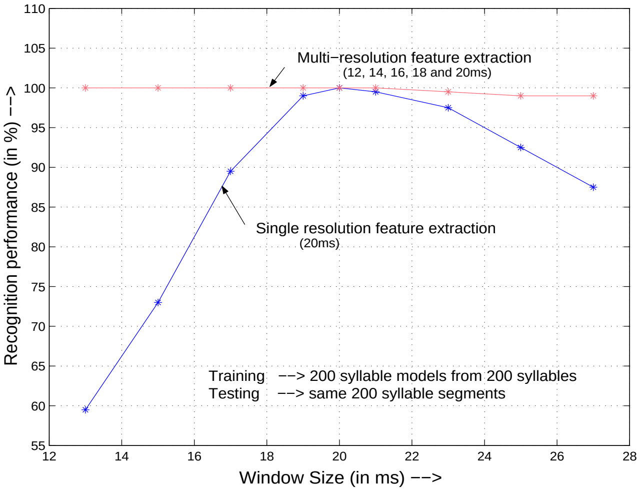

Fig. 1. Comparison between multiple resolution and single resolution feature extraction.

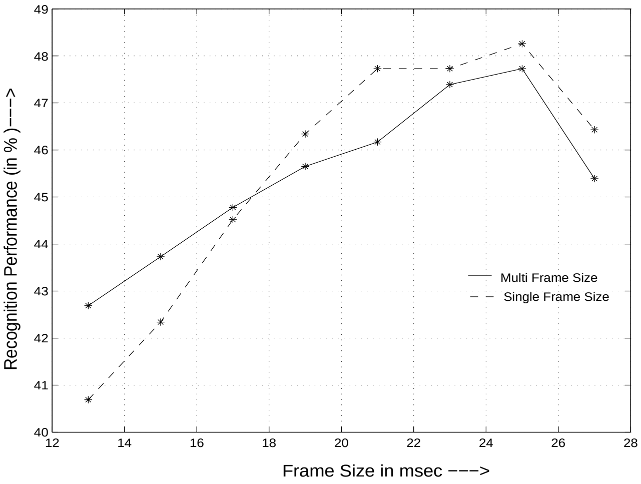

Fig. 2. Performance comparison of MFS and SFS feature extraction for female speaker

To analyse the performance of the recognition system for an entirely different speaker, a male data is also considered and the above described experiment is carried out. The results on male speaker's speech data is shown in Figure 3.

The recognition performance of the MFS feature extraction based models outperform the SFS feature extraction based models for the first few configurations (i.e., for the frame sizes 13, 15, and 17ms). But for the other cases, the performance of MFS models failed considerably when

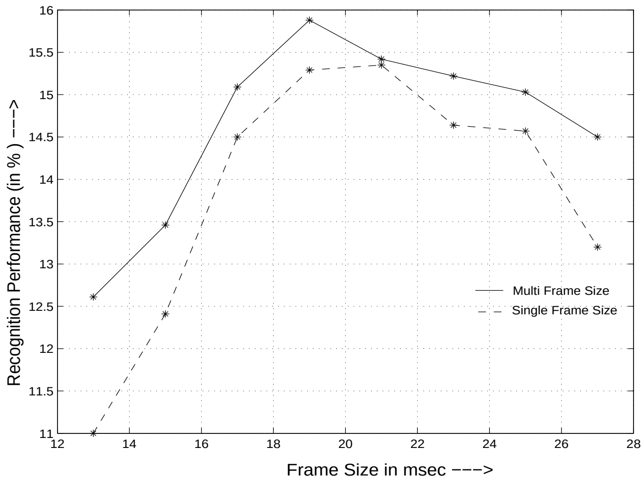

Fig. 3. Performance comparison between MFS and SFS feature extraction for male speaker

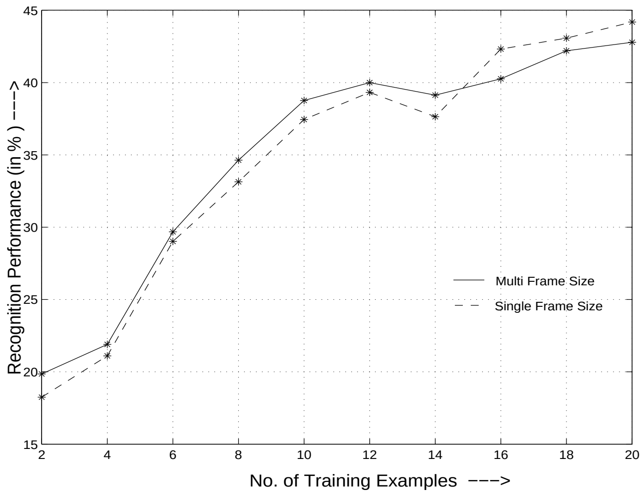

Fig. 4. Performance comparison of MFS and SFS models for different amount of training data.

compared with that of the SFS models. Further analysis on the performance shown that the MFS models generated with less number of examples outperformed SFS models and vice-versa. But, in case of male test speaker's speech data, the MFS models outperform for all the different configurations (see Figure 3).

In order to study the performance of the recognition system for different amount of training data, another experiment is carried out by gradually increasing the training data. The recognition performance of both MFS and SFS models are shown in the Figure 4. Interestingly, when the amount of training data is less (till 14 training examples), the performance of MFS models outperformed the SFS models. As the amount of training data is raised considerably, the performance of the SFS models starts dominating.

3. MFS AND MFR BASED FEATURE EXTRACTION

It is a known that speaking rate vary for different speakers. If the test speakers' speaking rate is different from that of trained speaker's speaking rate, the models generated with single frame rate may not be capable of handling such variation. By considering this issue, another experiment is carried out with multiple frame rates also. In order to derive advantage of both MFS and MFR based feature extraction, the models are generated with features extracted using MFS and MFR(with 5 different frame sizes and 5 different frame rates). The experiments are carried out in two different ways: (1) MFS and MFR based feature extraction during training alone and (2) MFS and MFR based feature extraction during training and testing, which are explained below.

3.1. MFS and MFR During Training alone

For a gender dependent speech recognition system, five female speaker's data is considered for training, The features are extracted using MFS and MFR based feature extraction and models (3 states, 1 mixture/state) are generated. During recognition, two different female and one male speaker's speech data are considered. In this experiment, during recognition, features are extracted with single frame size and single frame rate and the results are tabulated in Table 1.

Table 1. *Performance analysis of speech recognition system using MFS and MFR (during training alone) and SFS*

Speaker	Performance in %	
	MFS and MFR	SFS
Speaker I (female)	46.69	49.13
Speaker II (female)	44.53	47.58
Speaker III (male)	21.89	24.18

The speech recognition performance (refer Table 1) shows that, for all the speakers (male and female), the models generated using the single frame size and single frame rate based feature extraction performed better. To see the effect of MFS and MFR based feature extraction in the recognition phase, another experiment is carried out as explained in the next Subsection.

3.2. MFS and MFR During Training and Testing

In this experiment, the models are generated as explained in the previous experiment and the difference here is that, during recognition also the features are extracted with MFS and MFR. Therefore, for each frame size and frame rate based features, a separate recognition experiment will be performed. During recognition, the 3-best results are taken for each syllable and the final recognition is based on combining the results at the rank level. The 3-best results are taken for each syllable using different frame size and frame rate feature extraction (totally 10, 3-best results) and weights are assigned to the 3-best results based on their position(e.g 1 for first position, 0.8 for the second position and 0.6 for third position and so on) and ranks are computed for the 10, 3-best results. The syllable with highest rank is chosen as the recognized syllable.

The results for different test speakers' speech data (2 female and 1 male) is tabulated in Table 2. If we compare the performance of the recognition system in which MFS and MFR feature extraction is used with that of the recognition system in which single resolution is used, there is an observable improvement in the performance, i.e., 4% improvment in the case of female test data and 1.4% in the case of male data.

Table 2. *Performance analysis of speech recognition using MFS and MFR (with Ranking) and SFS*

Speaker	Performance in %	
	MFS and MFR	SFS
Speaker I (female)	53.3	49.13
Speaker II (female)	51.4	47.58
Speaker III (male)	25.6	24.18

4. FUTURE WORK AND CRITICISMS

The performance of the speech recognition system using MFS and MFR based feature extraction can be tested when the speech signal is corrupted by additive Gaussian noise to check the effect of its robustness for noisy speech data. The same approach can be applied for gender independent system also. A complete analysis should be made on how the multiple frame sizes and multiple frame rates are affecting the system's performance during recognition. Though there is an improvement in the performance of the recognition system, using of MFS and MFR based feature extraction during testing also, the computational complexity is considerably increased.

5. CONCLUSION

In this paper, an attempt is made to make use of multiple frame size and multiple frame rate based feature extraction for speech recognition task. The MFS based feature extraction is shown to perform better when the gender-dependent (female) models are tested with other gender (male) speech data. The effect of different amount of training data is analyzed and shown that MFS based feature extraction technique performs better than that of conventional feature extraction procedure. when the amount of training data is limited. Further, when MFS based feature extraction technique is combined with MFR based feature extraction, it is shown that a considerable improvement can be achieved over single frame size feature extraction procedure.

6. REFERENCES

[1] Macias-Guarasa, J., Ordonez, J, Montero, J., M., Ferreiros, J., Cordoba, R., and Haro, L., F., D, " Revisiting Scenarios and Methods for Variable Frame Rate Analysis in Automatic Speech Recognition", EUROSPEECH 2003 - Geneva.

[2] Philippe Le Cerf and Dirk Van Compernolle, "A New Variable Frame Rate Analysis Method for Speech Recognition", Signal Processing Letters IEEE, Volume: 1, Issue: 12, December 1994, Pages: 185-187.

[3] Qifeng Zhu and Abeer Alwan, "On The Use of Variable Frame Rate Analysis in Speech Recognition", Proceedings 2000, IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume: 3 , 5-9 June 2000, Pages:1783 - 1786.

[4] Nagarajan, T., "Implicit Systems For Spoken Language Identification", PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Chennai, India, 2004.

[5] Sarada, G., L., Hemalatha, N., Nagarajan, T., Murthy, H., A.,"Automatic Transcription of Continuous Speech using Unsupervised and Incremental Training", ICSLP-2004, October, Korea.

[6] Database for Indian Languages, India, speech and vision laboratory, IIT Madras, Chennai-2001.