# ON THE USE OF VARIABLE FRAME RATE ANALYSIS IN SPEECH RECOGNITION

*Qifeng Zhu and Abeer Alwan*

Department of Electrical Engineering, UCLA
Los Angeles 90095, USA
{qifeng, alwan}@icsl.ucla.edu

## ABSTRACT

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. In this paper, we propose a Variable Frame Rate (VFR) algorithm. The algorithm results in an increased number of frames for rapidly-changing segments with relatively high energy and less frames for steady-state segments. The current implementation used an average data rate which is less than 100 frames per second. For an isolated word recognition task, and using an HMM-based speech recognition system, the proposed technique results in significant improvements in recognition accuracy especially at low signal-to-noise ratios. The technique was evaluated with MFCC vectors and MFCC vectors with enhanced peak isolation [4].

## 1. INTRODUCTION

In most speech-processing systems, speech signals are first windowed into frames; frames are typically 20-30 ms in duration and the frame step size is 10 ms. This is especially true for HMM-based automatic speech recognition (ASR) systems. The justification for such a segmentation is that speech signals are non-stationary and exhibit quasi-stationary behavior at the shorter durations.

It is known, however, that certain acoustic attributes of the speech signal can be manifested at very short durations (see for example, [3]). Such attributes may be critical for the identification and discrimination of speech sounds.

In this study, we propose a variable frame-rate (VFR) approach for analyzing speech signals. The technique results in an increased number of frames when the spectral characteristics of the signal change significantly and less frames otherwise. The frame step size can be as low as 2.5 ms. The algorithm can be implemented such that the average data rate of the system is the same, less, or greater than the fixed data rate approach that is typically used in ASR systems.

For an isolated word, HMM-based, recognition task, the proposed technique results in significant improvements in recognition accuracy especially at low signal-to-noise ratios. The technique was evaluated with MFCC vectors and MFCC vectors with enhanced peak isolation [4].

The technique proposed in this paper differs from the VFR techniques in [2] and [5]. Both papers did not evaluate their systems in the presence of background noise. In addition, in [2], the focus was on using VFR to reduce the data rate of the system while keeping the frame step size at 10 ms, and thresholds were chosen in an ad hoc manner. In [5], a theoretically-motivated VFR system was proposed, but the evaluation was only done with a DTW recognition system and did not show improvement in recognition accuracy.

## 2. MOTIVATION AND PRELIMINARY EXPERIMENTS

In a previous study [1], we examined the acoustics and perception of the place-of-articulation for the highly confusable nasal consonants /m, n/ in pre-stressed syllable-initial position with the vowels /a, i, u/. The database was collected at UCLA and consists of speech tokens by 2 male and 2 female talkers with 8 repetitions per syllable (192 tokens in total). The sampling rate was 16 kHz. Perceptual experiments were conducted both in quiet, and in the presence of additive white Gaussian noise (AWGN) and speech-shaped noise. Results showed that formant transitions, in general, play a larger role in identifying place than the murmur. Specifically, perceptual thresholds were correlated with the duration and relative amplitudes of the formant, especially F2, transitions which in turn were vowel dependent. For example, the duration of the F2 transitions in /na/ syllables were the longest, as shown in Table 1, and with relatively high energy leading to very robust perception of the sound in noise. In addition, /ma/ syllables were robust even though F2 transition was short, but the amplitude of F2 relative to F1 was the largest of all syllables.

| ma | mi | mu | na | ni | nu |
|---|---|---|---|---|---|
| 19 | 20.8 | 16.6 | 57.5 | 19.3 | 12.9 |

**Table 1.** Average F2 transition in millisecond for different syllables. Measurements were done manually.

To compare human and machine nasal recognition, an experiment was conducted using the HMM-based ASR system from Entropics Inc. (HTK 2.0). Endpoint detection using energy and zero-crossing measures was used. Each HMM model had 6 states. Training was done with half of the utterances, and testing, with the other half. The feature vector used was the Mel-Frequency Cepstral Coefficients (MFCC) with first and second derivatives. The window (Hamming) length was 25 ms and the frame step size was 10 ms. The noise in all the experiments is additive speech shaped noise. If the system is trained and tested with clean data, high recognition accuracy is achieved (90 percent). If the system is trained with clean data and tested with noisy data, recognition scores deteriorate. For example, at a SNR

0-7803-6293-4/00/$10.00 ©2000 IEEE.

of 3 dB, the recognition accuracy is 52 percent; the corresponding confusion matrix is shown in Table 2. The /Ca/ syllables are the most robust in noise and we attribute that result to the more pronounced formant transitions for those syllables. We speculate that the deterioration in recognition accuracy for /Ci/ syllables, in particular, is attributed to their very short and weak formant transitions.

| | ma | mi | mu | na | ni | noo | Correct rate (%) |
|---|---|---|---|---|---|---|---|
| ma | 13 | 0 | 0 | 3 | 0 | 0 | 81.2 |
| mi | 0 | 0 | 0 | 2 | 6 | 8 | 0.0 |
| nu | 1 | 0 | 2 | 8 | 0 | 5 | 12.5 |
| na | 0 | 0 | 0 | 16 | 0 | 0 | 100 |
| ni | 0 | 1 | 0 | 1 | 8 | 6 | 50.0 |
| nu | 0 | 0 | 0 | 5 | 0 | 11 | 68.8 |

**Table 2.** Nasal recognition results. Trained with clean data, and tested with additive speech shaped noise at a SNR=3dB.

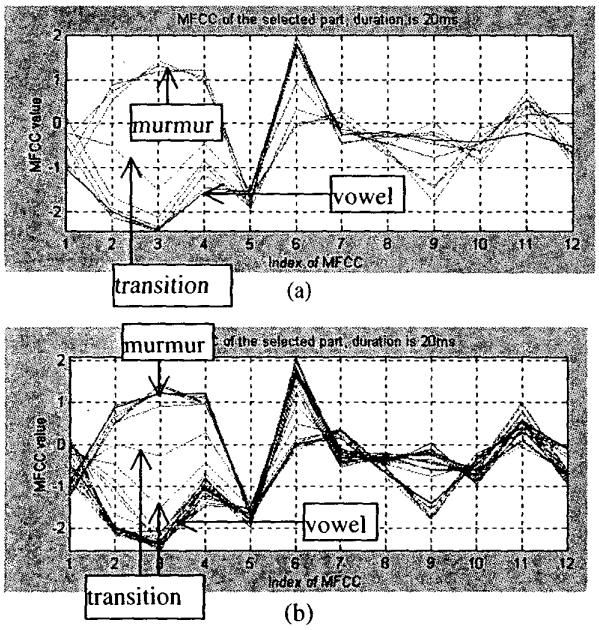


**Figure 1.** MFCC vectors around the transition of a /ma/ utterance. (a) Window step size = 10ms. (b) Window step size = 2.5ms.

A frame step size of 10 ms may not be sufficient to capture dynamic changes. To illustrate this point, Figure 1 shows plots of MFCC vectors along a 100 ms segment surrounding the formant transition region in a /ma/ syllable. The frame length is 20 ms, but the frame step size is 10 ms in (a) and 2.5 ms in (b). Speech is pre-emphasized and MFCC vectors are liftered. Note that the murmur and steady-state region of the vowel are represented by (perhaps an unnecessarily large) number of MFCC vectors, while the critical formant transition region (13 ms) is only represented by one vector with a 10 ms frame step size and 2 (distinct) vectors when the step size is reduced to 2.5 ms.

# 3. VARIABLE FRAME RATE (VFR) METHOD

## 3.1 The Algorithm

From the analysis described above, it is clear that computing frames every 10 ms is not adequate for representing rapidly changing segments although it is sufficient for representing relatively steady and long ones.

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, we propose a variable frame rate method in which the frame rate varies as a function of the spectral characteristics of the signal.

Using MFCC feature vectors, the variable frame rate algorithm is implemented as shown in Figure 2.

First, speech is analyzed with frame lengths of 25 ms (Hamming window) and a step size of 2.5 ms. We refer to these frames as the "dense frames". Second, the difference ($d(i)$, where $i$ is the time index,) between every two adjacent "dense frames" is calculated. The average of these differences is then calculated over the whole utterance. Third, based on the weighted differences, some frames are kept and others are discarded. In particular, "dense frames" around a formant transition will be kept, while at the steady part of the signal, frames will be picked sparsely.

It is important to note that the distance $d(i)$ is calculated as the *energy weighted* Euclidean MFCC distance: first the Euclidean distance of the MFCC vectors of two adjacent frames are calculated, then it is weighted by ($E$ - $\beta$), where $E$ is the log energy of that frame, and $\beta$ is a constant offset. This is different from the method proposed in [2] where the Euclidean MFCC distance was used. Energy weighting is important so that segments which exhibit changes but are low in energy are discarded, since they may not be noise robust. Our previous experiments have shown a clear relationship between the energy of formant transitions and perceptual noise robustness. In addition, our pilot ASR experiments using Euclidean MFCC distance did not yield high recognition accuracy in noise.

The two parameters $\alpha$, the threshold, and $\beta$, log energy offset, are chosen experimentally. The choice of $\alpha$ will determine the average data rate. For example, if $\alpha$ is 4 (ratio of the 10 ms step size and the dense step size of 2.5 ms), then the resulting total number of frames will be nearly the same as that in a front-end with a frame step size of 10 ms. If $\alpha$ is larger than 4, then the average data rate will be less than 100 frames per second and vice versa. In our implementation, $\alpha$ was chosen to be 6.8. The log energy offset $\beta$ was set to be the average $E$ (over the entire utterance) divided by 1.5.

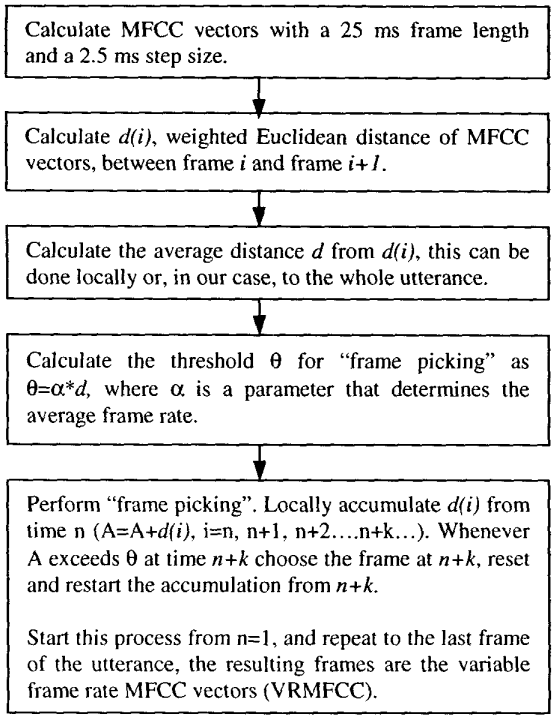


**Figure 2.** Flow chart of computing variable frame rate MFCC vectors.

## 3.2 An Example of VFR Analysis

Figure 3 illustrates how frames are picked for the utterance /ma/ as spoken by a male speaker. Part (a) shows a time waveform of the utterance. The upper part of (b) plots *d(i)*, the weighted feature distance between two adjacent frames - with a step size of 2.5 ms - and the lower part shows the result of the frame-picking algorithm where each bar indicates that a frame has been chosen for recognition. Note that near the transition region from the consonant to the vowel *d(i)* is large. For this example, 50 out of 200 dense frames are picked. Around the transition region, all the dense frames (spaced by 2.5 ms) are kept while in the steady-state part of the vowel, only 3-4 frames, out of 20 frames, are selected corresponding to a step size which is larger than 10ms.

## 3.3 Recognition with the VFR Front End

The variable frame rate method was used in ASR experiments using the nasal database described in Section 2, and the TIDIGITS database. In the experiments, the performance of the recognition system with two feature vectors were compared: MFCC, and MFCC vectors with peak enhancement [4] (hereafter referred to as MFCCP). First and second derivatives of these features were used. Training was done using clean data while testing was done with either clean or noisy data.

Results for the nasal recognition experiment are shown in Table 3. Clearly, the variable frame rate approach together with a method for enhancing spectral peaks, gives the best performance at low SNRs.

The VFR method was also used with the database "Studio Quality Speaker Independent Connected Digit Corpus" (TIDIGITS). Each left to right digit HMM model had 4 states, 2 mixtures, and a diagonal covariance matrix. 80 utterances from 80 speakers, (40 male and 40 female) were used to train each model. Test data were from the other 32 speakers (half male and half female).

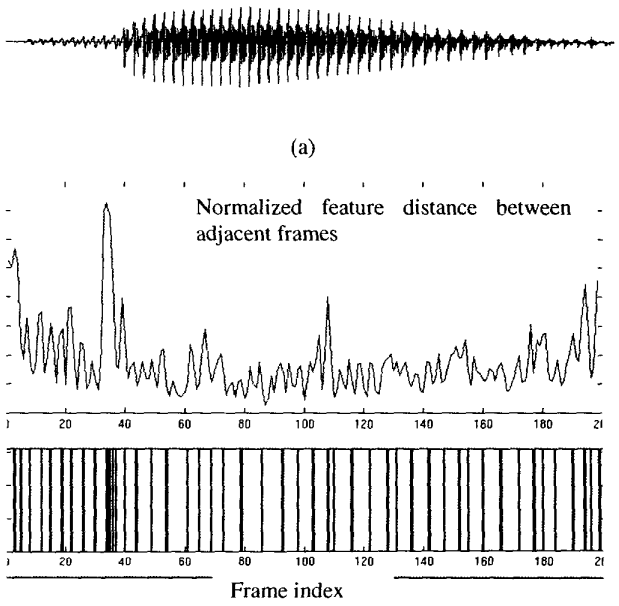


**Figure 3.** (a) The waveform of a /ma/ utterance. (b) The upper panel is the normalized *d(i)* distance, and the lower panel shows which frames were kept.

| | Clean | SNR=15dB | 5dB | 0dB |
|---|---|---|---|---|
| MFCC | 90 | 89 | 68 | 34 |
| MFCCP | 96 | 91 | 77 | 68 |
| VRMFCCP | 100 | 96 | 81 | 71 |

**Table 3.** Percent correct rates from different front-ends using the nasal database.

We compared MFCC and MFCCP with their variable frame rate versions. The results are shown in Figures 4 and 5 and is summarized in Table 4. The results clearly illustrate that applying the VFR method to each feature vector improves recognition performance especially at low SNRs. Increasing time resolution for rapidly changing segments, while keeping the time resolution low for steady parts, results in improved robustness.

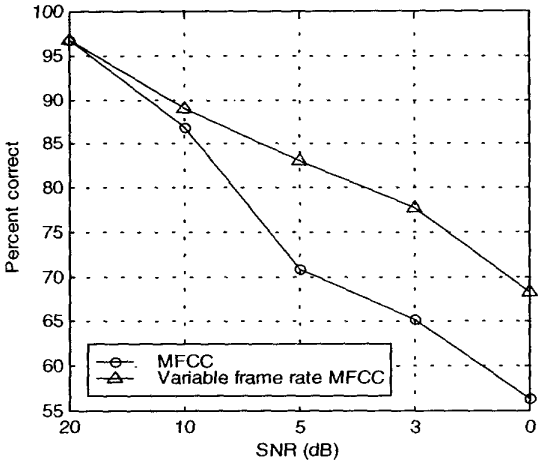


**Figure 4.** Recognition results expressed by word percent correct for MFCC and variable frame rate MFCC using the TIDIGITS database.

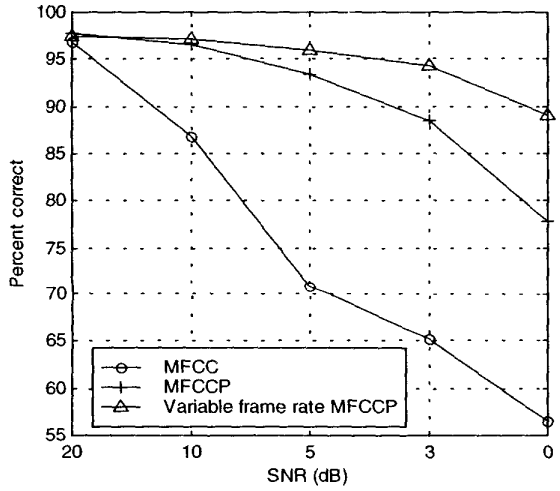


**Figure 5.** Recognition results expressed by word percent correct for MFCC, MFCC with peak isolation (MFCCP), and its variant frame rate version (VRMFCCP) using the TIDIGITS database.

| Percent correct | SNR=20 dB | 10dB | 5dB | 3dB | 0dB |
|---|---|---|---|---|---|
| MFCC | 96.87 | 86.83 | 70.85 | 65.20 | 56.43 |
| MFCCP | 97..81 | 96.55 | 93.42 | 88.40 | 77.74 |
| VRMFCCP | 97.49 | 97.18 | 95.92 | 94.36 | 89.03 |

**Table 4.** Recognition results summary for MFCC, MFCCP and VRMFCCP front ends using the TIDIGITS database.

# 4. SUMMARY AND CONCLUSION

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 ASR systems, is not sufficient to capture such dynamic changes. In this paper, we propose a Variable Frame Rate (VFR) algorithm. The algorithm results in an increased number of frames for rapidly-changing segments with relatively high energy and less frames for steady-state segments. It can be implemented such that the average data rate is the same, less, or more than a fixed 100 frames per second data rate. The current implementation used an average data rate which is less than 100 frames per second. For an isolated word recognition task, and using an HMM speech recognition system, the proposed technique results in significant improvements in recognition accuracy especially at low signal-to-noise ratios. The technique was evaluated with MFCC vectors and MFCC vectors with enhanced peak isolation [4].

The novel properties of the proposed VFR algorithm are 1) using energy weighted MFCC distance, 2) allowing a frame step size as low as 2.5 ms, and 3) a novel method for frame selection.

**Acknowledgement**

Work is supported in part by NSF and NIH.

# 5. REFERENCES

[1] Abeer Alwan, Jeff Lo, and Qifeng Zhu "Human and Machine Recognition of Nasal Consonants in Noise", Proceedings of the 14th International Congress of Phonetic Sciences, Vol. 1, p. 167-170, 1999.

[2] 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.5, (no.2), April 1991. p.169-79.

[3] Kenneth Stevens, *Acoustic Phonetics*, MIT Press, 1998.

[4] Strope, B. and Alwan, A. 1997. "A model of dynamic auditory perception and its application to robust word recognition", IEEE Trans. on Speech and Audio Processing, Vol. 5, No. 5, p. 451-464.

[5] Young, S.J. and Rainton, D. "Optimal frame rate analysis for speech recognition." IEE Colloquium on Techniques for Speech Processing (Digest No.181), London, UK. 17 Dec. 1990, p.5/1-3.