ON THE USE OF VARIABLE FRAME RATE ANALYSIS IN SPEECH RECOGNITION

Author : Qifeng Zhu and Abeer Alwan

Professor:陳嘉平
Reporter:葉佳璋

Outline

- Introduction
- Motivation and preliminary experiments
- Variable frame rate (VFR) method
- Summary and conclusion

Introduction

- In most of speech-processing system, speech signal are first windowed into frames;

- In the study, propose a variable frame-rate approach for analyzing speech signal.

Introduction

- The technique was evaluate with MFCC vector and MFCC vector with enhance peak isolation.

- The proposed technique results in significant improvements in recognition accuracy especially at low signal-to-noise ratios.

Motivation and preliminary experiments

- The database was collected at UCLA and consists of speech token by 2 male and 2 female talkers with 8 repetitions per syllable(192 tokens in total).

- The sampling rate was 16kHz.

- An experiment was conducted using the HMM-based ASR system from Entropics Inc(HTK2.0). Each HMM model had 6 states.

Preliminary experiments result

- Result showed that formant transitions play a larger role in indentifying place.

- Specifically, perceptual thresholds were correlated with the duration and relative amplitudes of the formant, especially F2, transitions which in turn were vowel dependent.

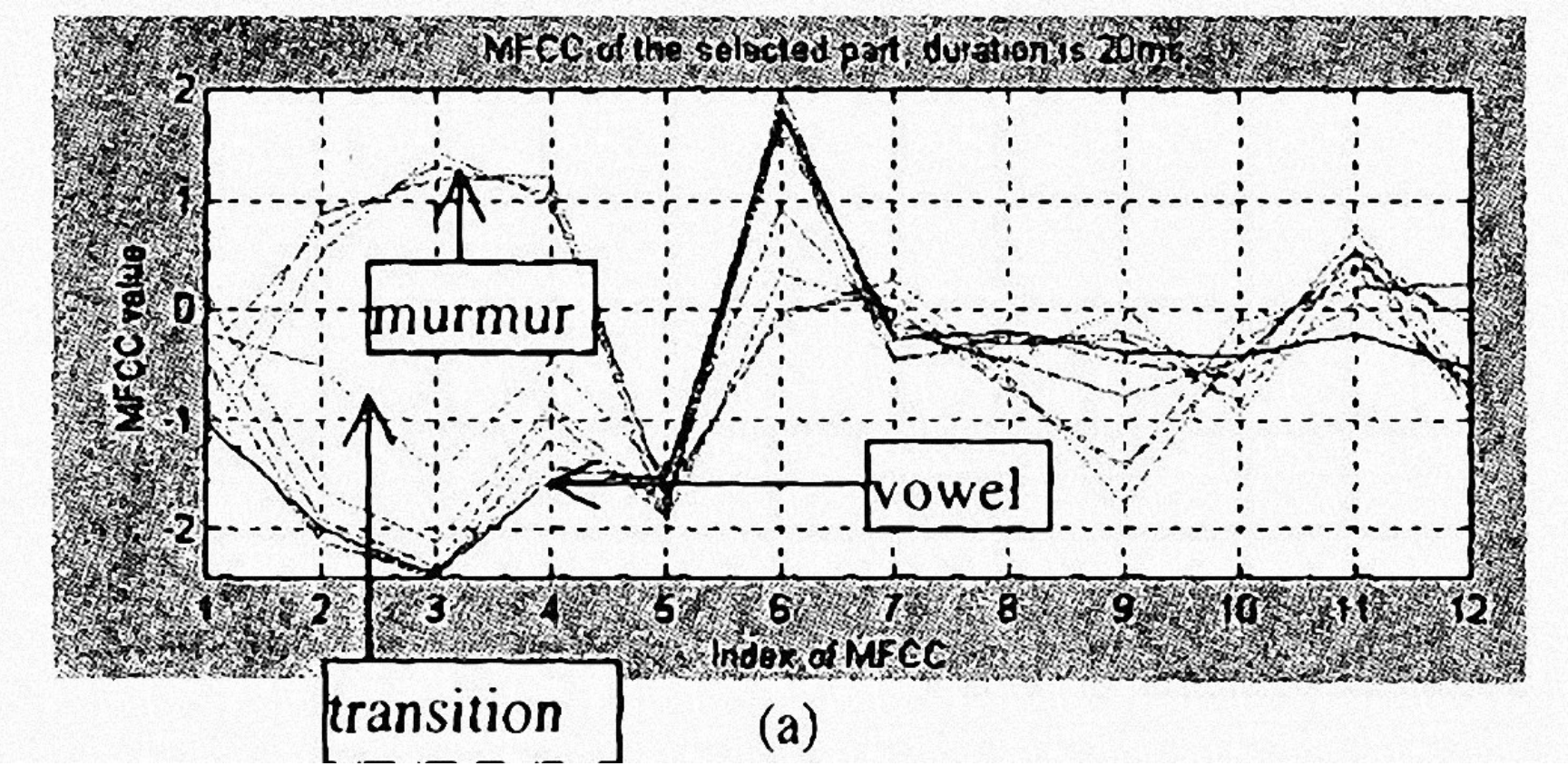
Preliminary experiments result

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

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

	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.


MFCC of the selected part, duration is 20ms
MFCC value
2
1
0
-1
-2
murmur
vowel
transition
1
2
3
4
5
6
7
8
9
10
11
12
Index of MFCC
(a)

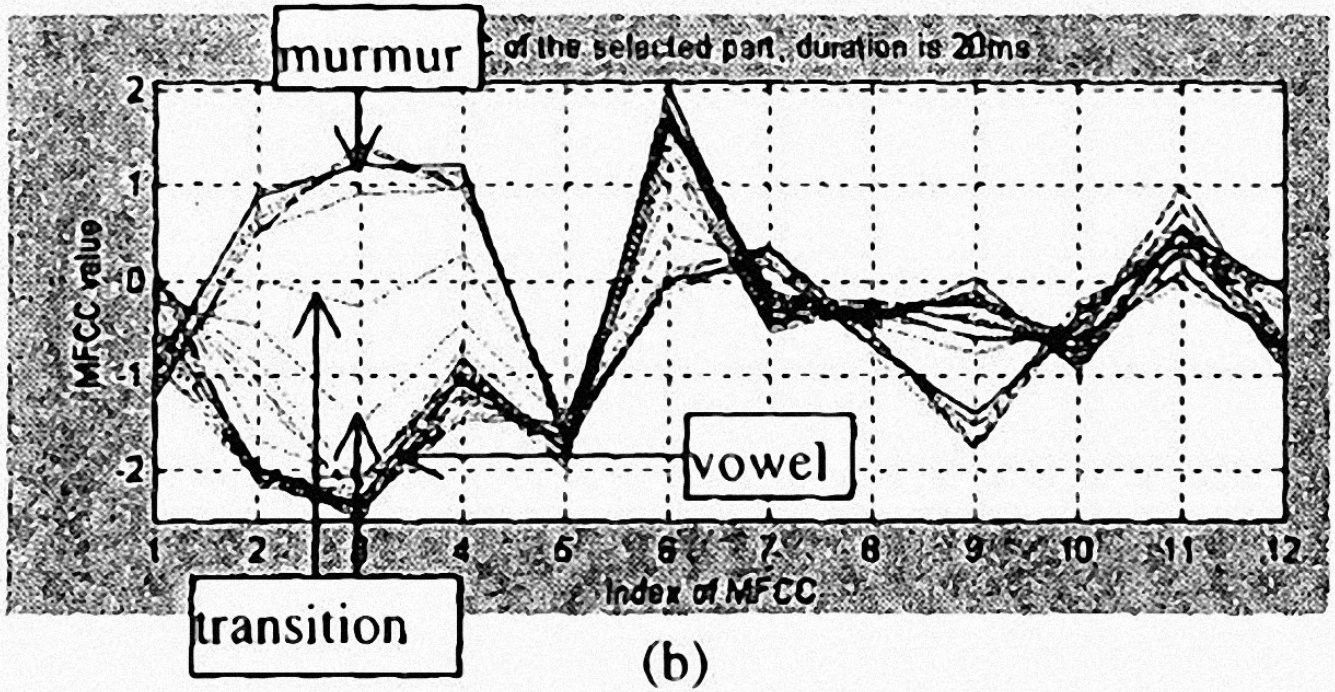


Figure 1. MFCC vectors around the transition of a /ma/ utterance. (a) Window step size = 10ms. (b) Window step size = 2.5ms.

Variable frame rate (VFR) method

- From the analysis describe above, it is clear that computing frames every 10 ms is not adequate for representing rapidly changing segment although it is sufficient for representing relatively steady and long one.

- One solution to this problem is increasing the frame rate, but this would unnecessarily increase the computational load of ASR system and is not needed for steady segment.

The algorithm

- First, speech is analyzed with frame length 25 ms(Hamming window) and a step size of 2.5ms. We refer to these frames as the “dense frame”.

- Second, the difference(d(i), where i is the time index) between two adjacent ”dense frame” is calculated. The average of these difference is then calculated over the whole utterance.

- Third, based on the weighted differences, some frames are kept and other are discard. In particular, “dense frames” around a formant transition will be kept, while at the steady part of the signal, frames will be picked sparsely.

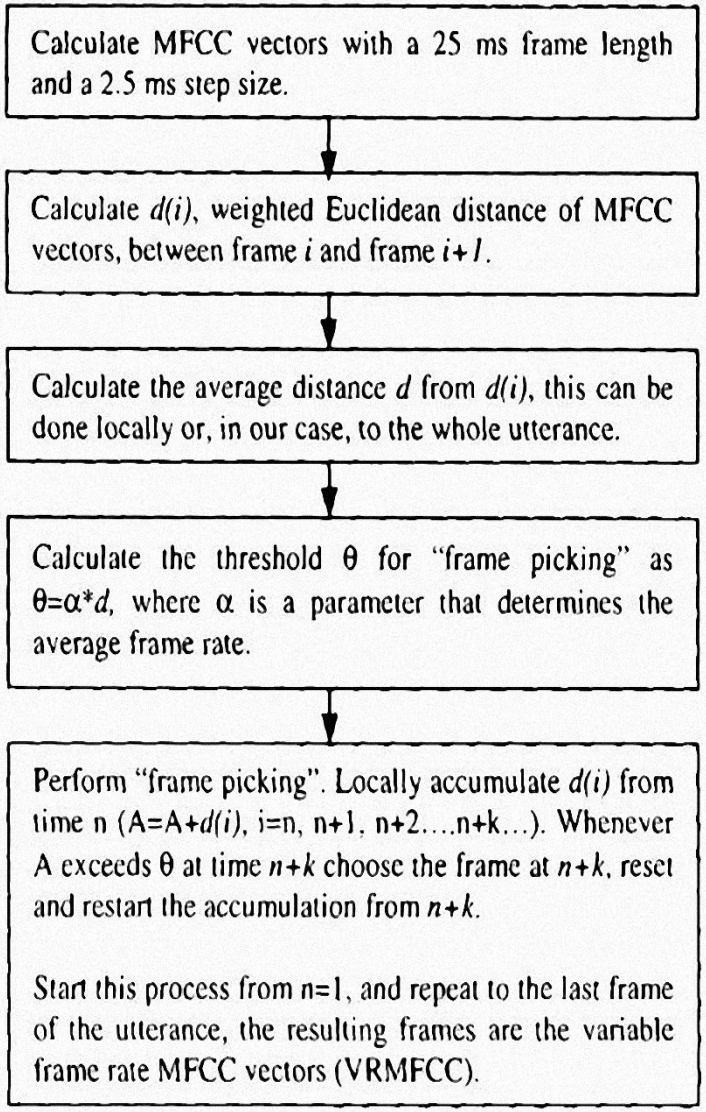


Figure 2. Flow chart of computing variable frame rate MFCC

The algorithm(cont.)

- It is important to note the distance d(i) is calculated as the energy weighted Euclidean MFCC distance: first the MFCC vectors of two adjacent frames are calculated, then it is weighted by (E-ß), where E is the log energy of that frame, and ß is constant offset.

- Energy weighting is important so that segments which exhibit changes but are low in energy are discarded, since they may not be noise robust.

The algorithm(cont.)

- The two parameters α, the threshold, and ß log energy offset, are chosen experimentally. The choice of α will determine the average data rate.

- For example if α is 4 then the resulting total number of frames will be nearly the same as that in a front end with frame step size of 10 ms. If α is larger then 4, then the average data rate will be less than 100 frames.

- In this implementation, α is chosen to be 6.8 the log energy offset ß was set to be the average E(every entire utterance) divided by 1.5.

An example of an VFR analysis

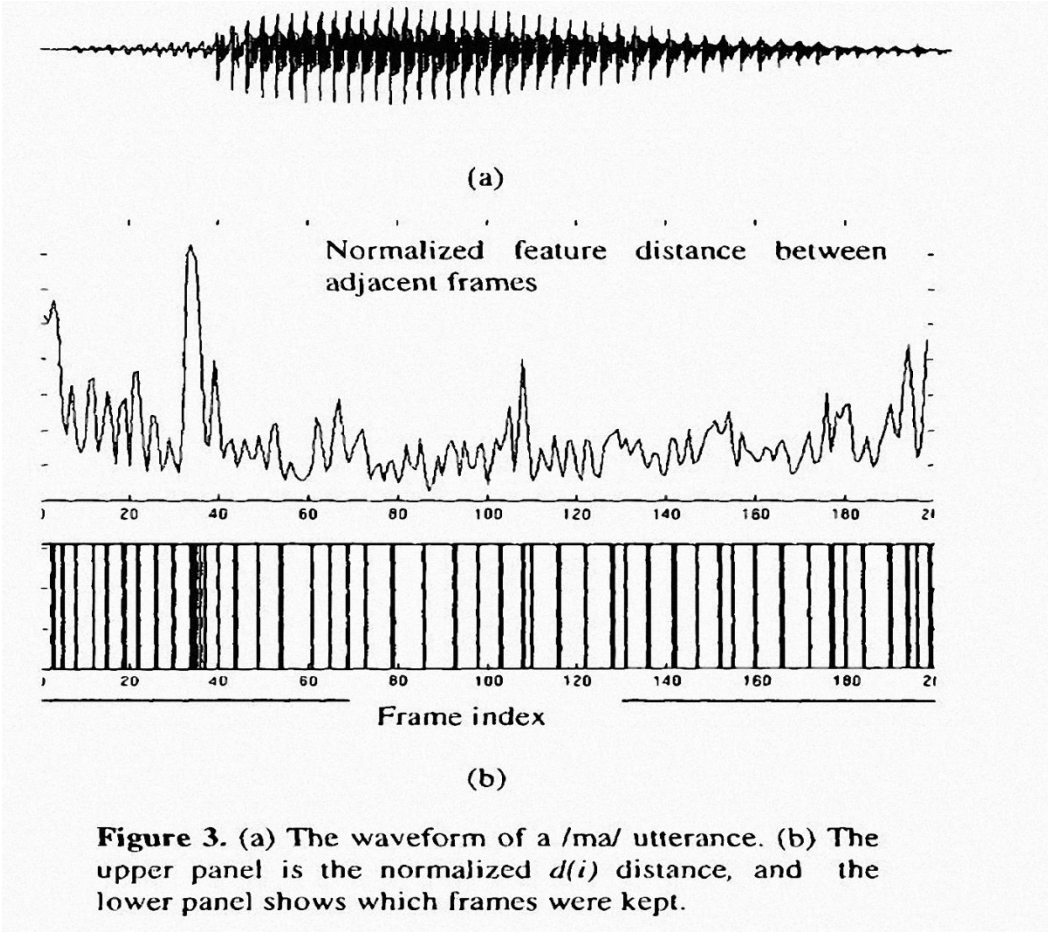


Figure 3. (a) The waveform of a /ma/ utterance. (b) The upper panel is the normalized *d(i)* distance, and the lower panel shows which frames were kept.

Peak isolation

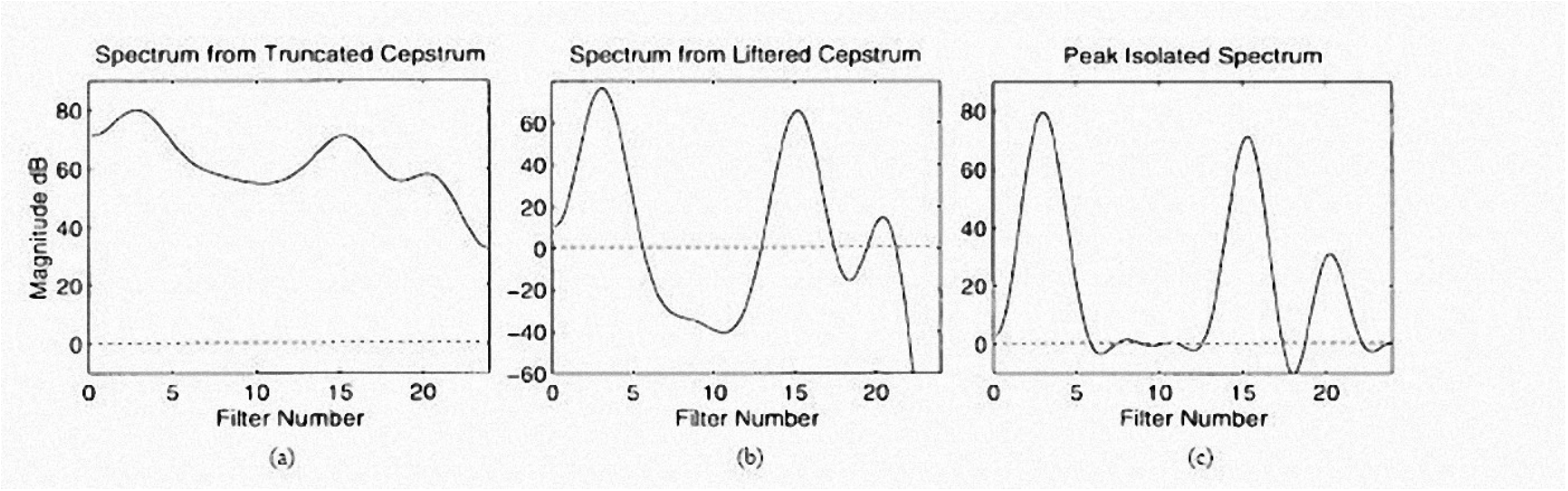
- Both speech perception and the response of individual auditory nerves are extremely sensitive to the frequency position of local spectral peak.

- Similarly perceptual discrimination of vowels is more sensitive to the frequency location of spectral peaks than to other aspects of the spectral shape. It suggest that the auditory system may derive a noise-robust representation by attending to the frequency peaks.

Peak isolation(cont.)

- Weighting the cepstral vector specifies the relative emphasis of different types of log-spectrum variations. A raised-sine lifter deemphasizes slow changes with frequency, often associate with overall level and vocal driving-function characteristics.

- The valleys are removed by half-wave rectifying the log spectral estimate implied after raised-sin liftering, and a final vector is obtained by transforming back to the cepstral domain.


Spectrum from Truncated Cepstrum
Magnitude dB
80
60
40
20
0
0 5 10 15 20
Filter Number
(a)
Spectrum from Liftered Cepstrum
60
40
20
0
-20
-40
-60
0 5 10 15 20
Filter Number
(b)
Peak Isolated Spectrum
80
60
40
20
0
0 5 10 15 20
Filter Number
(c)

Recognition with VFR Front End

- The variable frame rate was used in ASR experiments using the nasal database described before, and the TIDIGITS data base.

- The performance of the recognition system with two feature vector were compared: MFCC, and MFCC vector with peak enhanced(MFCCP).

- The VRF method was also used with TIDIGITS.

	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.

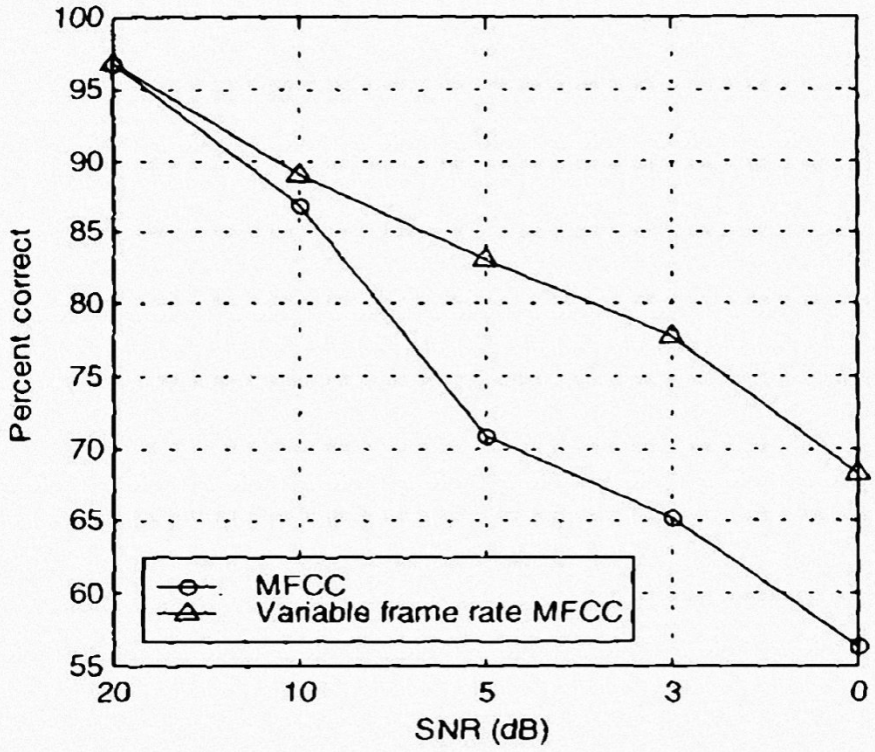


Figure 4. Recognition results expressed by word percent correct for MFCC and variable frame rate MFCC using the TIDIGITS database.

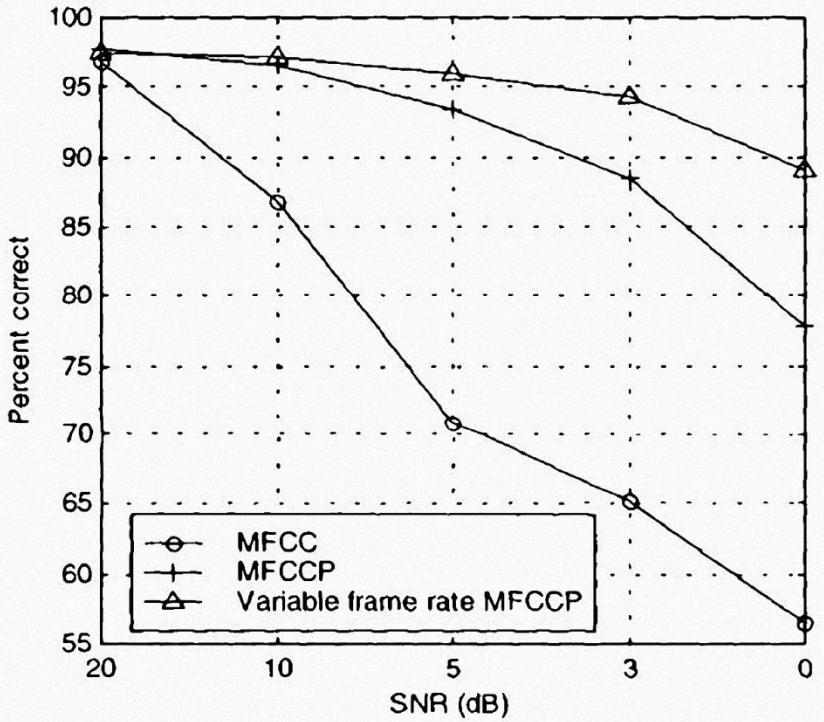


Figure 5. Recognition results expressed by word percent correct for MFCC, MFCC with peak isolation (MFCCP), and its variant frame rate version (VRMFCCP) using the TIDIGITS database.

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.

Summary Conclusion

- Changes in spectral characteristics are import cues for discriminating and identifying speech sounds.
- The novel properties of the propose VRF algorithm are

  1.using energy weighted MFCC distance.

  2.allowing a frame step size as low as 2.5 ms.

  3.a novels method for frame selection.