EFFECT OF DIFFERENT SAMPLING RATES AND FEATURE VECTOR SIZES ON SPEECH RECOGNITION PERFORMANCE

Conrad Sanderson and Kuldip K. Paliwal
School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
C.Sanderson, K.Paliwal@me.gu.edu.au

ABSTRACT

In this paper, we conduct a systematic study to evaluate the effect of sampling rate and feature-vector size on the performance of a Hidden Markov Model (HMM) based speech recognizer. We investigate the use of the following two types of features: Linear Prediction (LP) derived Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) [1, 2, 3]. We demonstrate that for the LPCC front-end, the optimum sampling rate and feature-vector size are 12 kHz and 14, respectively. We also show that for different sampling rates, accuracy peaks at different sizes of the feature-vector. For the MFCC front-end, the optimum feature-vector size and sampling rate are 14 and 14 kHz, respectively.

1. INTRODUCTION

Speech recognition systems reported in the literature use different sampling rates and feature-vector sizes. The effect of sampling rate and feature-vector size on the recognition performance has not been studied - most researchers use ad hoc values for the sampling rate and feature-vector size.

A lower sampling rate would reduce the storage requirements for a database, while a smaller feature-vector size would reduce training and recognition time. The problem is then to find a combination of sampling rate and feature-vector size which maximises recognition performance.

We performed a systematic study to evaluate the effect of sampling rate and feature-vector size on the performance of a HMM-based speech recognizer, using speaker- and context-independent phoneme models. 66 speech recognition experiments were done where the sampling rate was varied from 6 to 16 kHz in steps of 2 kHz. The LPCC and MFCC front-ends were used in the experiments.

In the following section, we briefly describe the database, pre-processing, and the training and testing methods used in our experiments.

2. DATABASE

A subset of the TIMIT database was used for training and testing. Only SX sentences were used during training (2310 sentences) and the core test set (192 sentences) was used for testing in order to minimize the amount of time taken. Downsampled versions (at 14, 12, 10, 8 and 6 kHz) of the original 16 kHz speech files were created.

3. PRE-PROCESSING

The speech files were pre-processed on a frame by frame basis, with a frame length of 20 ms and frame shift of 10 ms. For each frame, a Hamming window was applied before Linear Prediction Cepstral coefficients (LPCC) or Mel Frequency Cepstral coefficents (MFCC) were calculated. 20 bins were used for Mel Frequency analysis. The feature-vector was made up of the cepstral coefficients, normalised energy and their corresponding deltas. The figures presented in this paper only show the number of primary features in each vector (i.e., excluding energy and deltas). The actual size of the feature-vector is: *2*(size of primary vector)+2*.

4. HMM SPEECH RECOGNIZER

The HTK v2.02 (HMM Toolkit) package was used to train and test 48 context-independent, 4-mixture HMMs. An overview of HTK can be found in [3].

The results of a recognition experiment were mapped to a set of 39 phones, as described in [4] for performance evaluation.

In the figures presented in this paper, *% correct* is equal to $H/N*100\%$, and *% accuracy* is $(H-I)/N*100\%$, where $H =$ *Hits (number of phones correctly recognized)*, $N =$ *total number of phones* and $I =$ *number of incorrect phones that were inserted by the recognizer.*

For training of each HMM, 10 iterations were used for the initialisation as well as the re-estimation, while 2 passes of embedded re-estimation were performed.

5. EXPERIMENTS AND RESULTS

Figures 1 and 2 show the recognition performance for the LPCC front-end for six different primary feature-vector sizes, ranging from 8 to 18, against the sampling rate, which ranges from 6 to 16 kHz. Figure 1 shows feature-vector sizes 8 to 12 while Fig. 2 shows sizes 14 to 18.

Figures 3 and 4 contain the same data as Figs. 1 and 2, however the data are represented differently - the figures show recognition performance at different sampling rates against the size of the primary feature-vector, ranging from 8 to 18.

In Figs. 1 and 2 it can be seen that accuracy peaks at different sampling rates for different sizes of the primary feature-vector. Moreover, the accuracy drops as the sampling rate increases.

In Fig. 1, the maximum recognition accuracy rate seems to move from 10 to 12 kHz sampling frequency as the size of the feature-vector increases. In Fig. 2, the maximum accuracy occurs at 10 kHz for the vector sizes of 16 and 18 and at 12 kHz for the vector size of 14. Note that at 14 and 16 kHz we had trouble training a small number of HMMs - this affected the results obtained.

In Figs 3 and 4 it can be seen that generally as the feature-vector increases in size, so does the accuracy. For

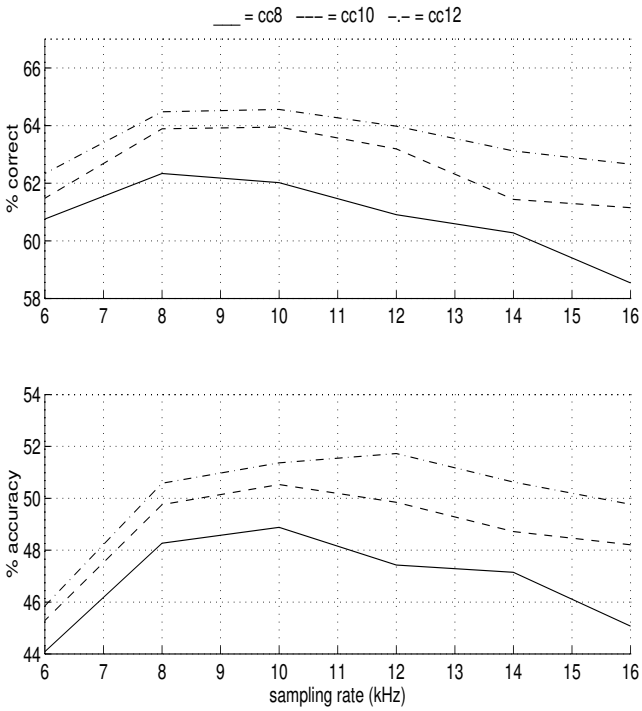

Figure 1. Recognition performance at different sampling rates for the LPCC front-end. Three sizes of primary feature-vector are shown: 8, 10 and 12.

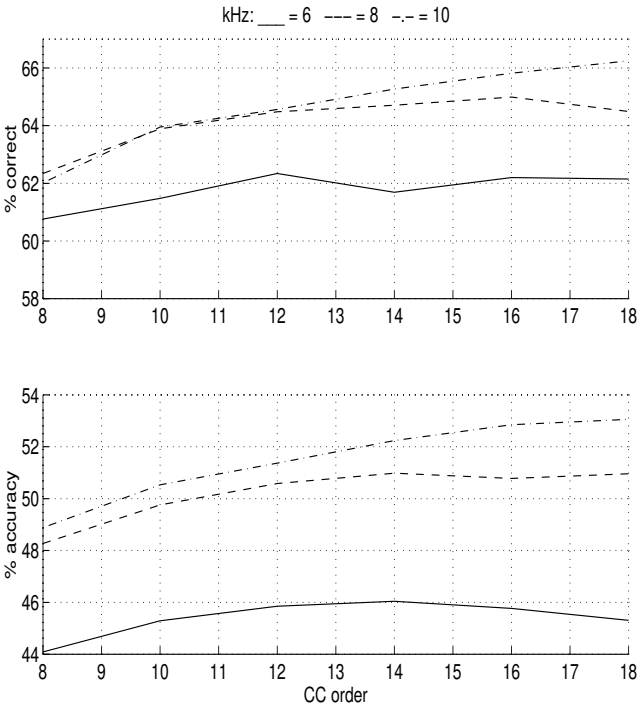

Figure 3. Recognition performance at different sizes of primary feature-vectors for the LPCC front-end. Three sampling rates are shown: 6, 8 and 10 kHz.

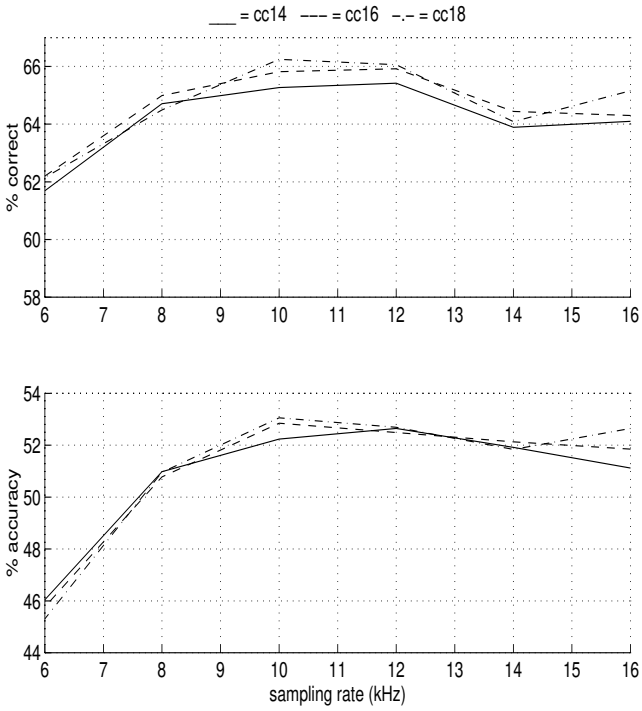

Figure 2. Recognition performance at different sampling rates for the LPCC front-end. Three sizes of primary feature-vector are shown: 14, 16 and 18.

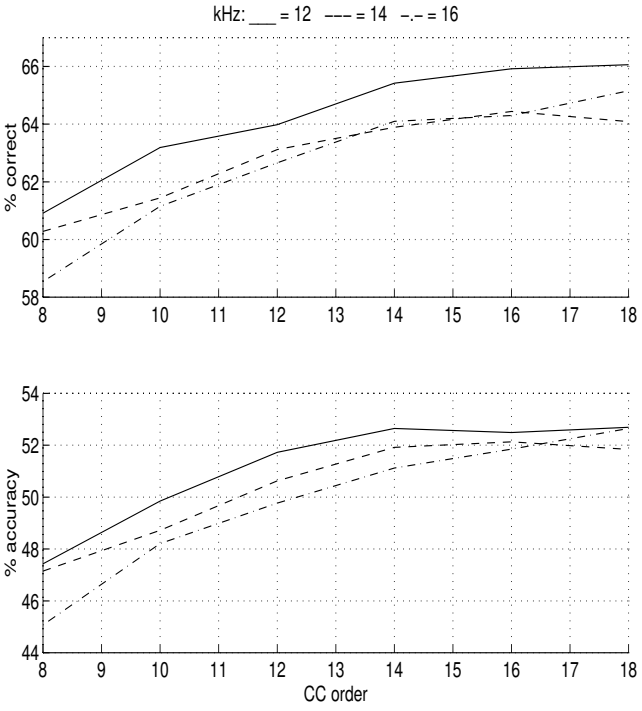

Figure 4. Recognition performance at different sizes of primary feature-vectors for the LPCC front-end. Three sampling rates are shown: 12, 14 and 16 kHz.

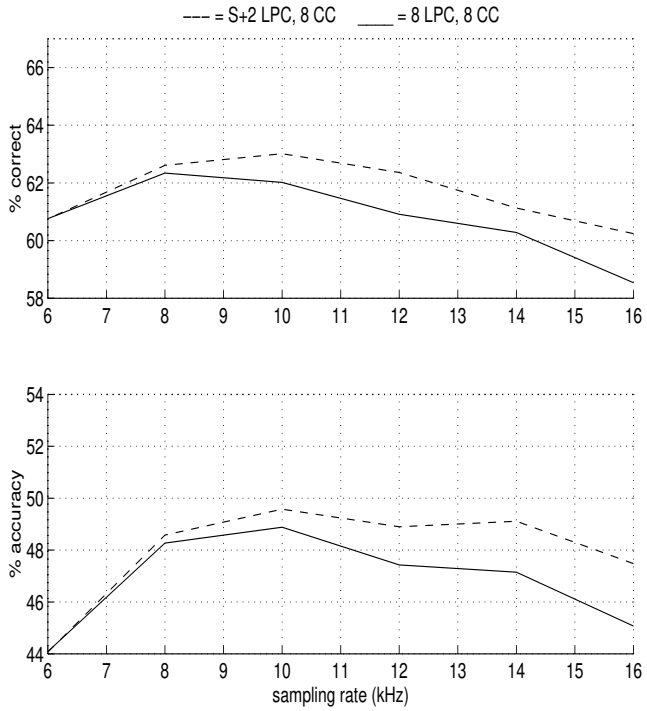

Figure 5. Recognition performance for LPCC with constant size of the primary features derived from a varying number of Linear Prediction coeffcients. S+2 denotes the number of LP coefficients: sampling rate/kHz + 2

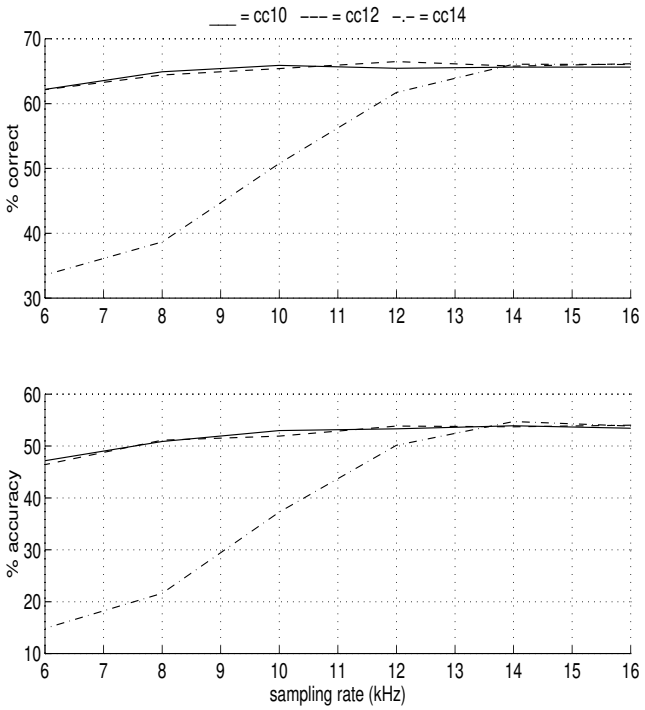

Figure 6. Recognition performance for MFCC with varying sizes of the primary features

a 6 kHz sampling rate, the accuracy peaks at the feature-vector size of 14, then decreases as the vector size increases. For the case of 8 kHz sampling rate, the accuracy levels off at vector size of 14, indicating that extra features have very little contribution. Figure 4 indicates that the optimum sampling rate and feature-vector size combination is 12 kHz and 14 respectively.

In Fig. 5 we investigated the effect of using a constant number of cepstral coefficients while the number of LP coefficients as well as the sampling rate varied. We tested the "rule of thumb" [5] which states that the order of LP analysis should be that of the sampling rate (in kHz) + 2. We compared this to the case where the order of LP remained constant. It can be seen that using LP coefficients from which a lower number of cepstral coefficients are generated increases both accuracy and correct rate.

Figure 6 shows the recognition performance for the MFCC front-end for three different feature-vector sizes at sampling rates from 6 to 16 kHz. It can be seen that for all feature-vector sizes an increase in the sampling rate generally increases accuracy. For the feature-vector sizes of 10 and 12, the correct and accuracy rates decrease gracefully as the sampling rate is decreased. However for the vector size of 14, the recognition performance drops significantly for sampling rates below 14 kHz. This is most likely due to the number of bins used during Mel Frequency analysis being too high for the low sampling rates.

For feature-vector sizes of 10 and 12, the accuracy levels off at a sampling rate of 12 kHz. For the case of feature-vector size of 14, accuracy levels off at a sampling rate of 14 kHz. The figure also implies that for the MFCC front-end, the optimum feature-vector size and sampling rate combination is 14 and 14 kHz respectively.

Figures 7 and 8 contain the same data as Fig. 6, but present recognition performance at different sampling rates against the primary feature-vector ranging from 10 to 14.

In Fig. 7, it can be seen that increasing the feature-vector size to a value greater than 12 reduces accuracy for sampling rates of 6, 8 and 10 kHz.

Figure 8 shows that accuracy stays relatively constant for sampling rates of 14 and 16 kHz for all feature-vector sizes, while for the sampling rate of 12 kHz the accuracy decreases slightly as the feature-vector size is increased from 12 to 14.

6. CONCLUSION

We have demonstrated that for a speech recognition system, based on HMMs and using the LPCC front-end, accuracy peaks at different sampling rates for different sizes of the feature-vector. This indicates that the selection of a sampling rate and feature vector size significantly affects the performance of the recognizer.

Using the LPCC front-end, the optimum sampling rate and feature-vector size combination for the recognizer used in our experiments is 12 kHz and 14, respectively. For the MFCC front-end, the optimum feature-vector size and sampling rate combination is 14 and 14 kHz, respectively.

REFERENCES

[1] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. 28, No. 4, pp. 357-366, Aug. 1980.

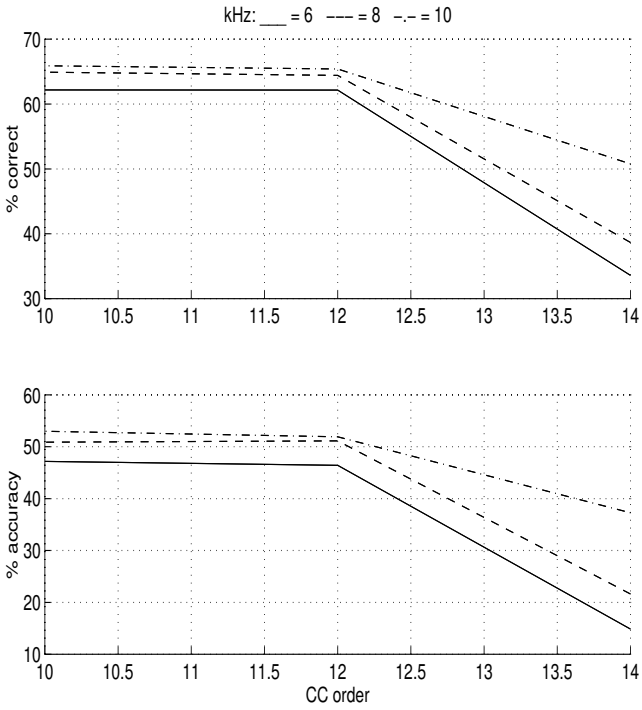

Figure 7. Recognition performance at different sizes of primary feature-vectors for the MFCC front-end. Three sampling rates are shown: 6, 8 and 10 kHz.

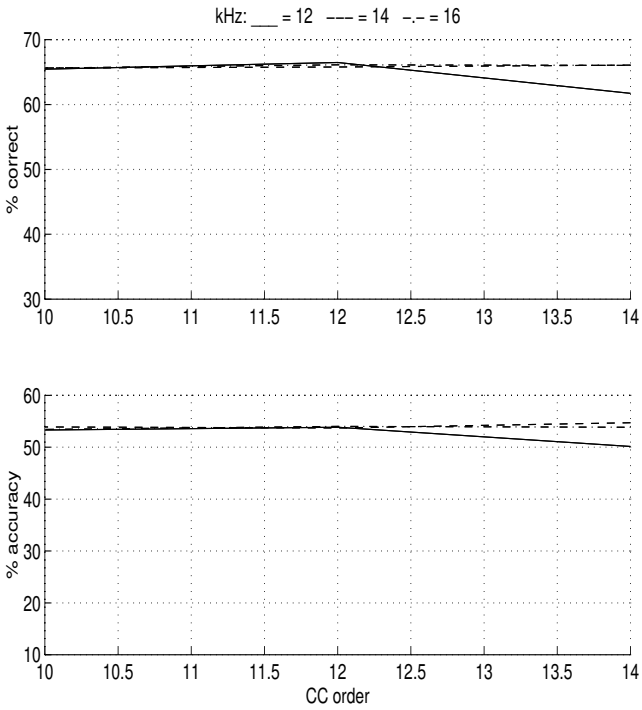

Figure 8. Recognition performance at different sizes of primary feature-vectors for the MFCC front-end. Three sampling rates are shown: 12, 14 and 16 kHz.

[3] S. Young, "A Review of large-vocabulary continuous-speech recognition", *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pp 45-57, Sep. 1996.

[4] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden Markov models", *IEEE Trans. Speech and Audio Processing*, Vol. 37, No. 11, pp. 1641-1648, Nov. 1989.

[5] K. K. Paliwal, "Speech processing techniques", *Advances in Speech, Hearing and Language Processing*, Vol. 1, pp. 1-78, 1990.