EFFECT OF SPEECH CODERS ON SPEECH RECOGNITION PERFORMANCE

B.T. Lilly and K.K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia

ABSTRACT

Speech coders with bitrates as low as 2.4 kbits/s are now being developed for speech transmission in the telecommunications industry. For speech coders to work at this reduced bitrate, some speech information has to be removed and it is only natural to expect that the performance of speech recognition systems will deteriorate when coded speech is applied as input to a recognition system. In this paper, the results of a study to examine the effects speech coders have on speech recogntion are presented. Six different speech coders ranging from 4.8 kbits/s to 40 kbits/s are used with two different speech recognition systems 1) isolated word recogntion and 2) phoneme recogntion from continuous speech. The effects on speech recognition performance by tandeming each of the speech coders are also presented.

1. INTRODUCTION

Significant advances have been made in the area of speech coding over the last 15 years and speech coding algorithms are now available which can produce communication quality speech at a bitrate as low as 2.4 kbits/s [1]. These advances combined with current DSP hardware technology have made it possible to utilize speech coding in telecommunication applications. This can be evidenced from the development of a number of speech coding standards [1], which span the bitrate from 4.8 kbits/s to 64 kbits/s.

Currently speech and speaker recognition systems work on speech digitized using linear PCM. In the future, speech and speaker recognition systems will be used from a remote location which means that they will operate on a speech signal which will have gone through an unknown number of speech coders during its transmission. Since speech coders introduce distortion into the speech signal, it is only natural to expect that the recognition performance of these systems will deteriorate with the reduction in the bitrate [2]. However, since most of the low bitrate speech coders are optimized for some perceptually-related criterion [1], it is not possible to predict their effect on the recognition performance.

In this paper, the influence of speech coders on speech recognition performance is studied using two experiments. Firstly using both isolated word and phoneme based speech recognition systems, we examine how the distortion introduced by speech coders at different bitrates affects the recognition performance. The second set of experiments examines how tandeming the speech coders affects the performance of the two recognition systems.

2. SPEECH CODERS

In order to digitize telephone-bandwidth speech, a typical linear PCM system uses a sampling rate of 8 kHz and a resolution of 16 bits/sample. This means that digitized speech has a bitrate of 128 kbits/s. Obviously this bitrate is too large for speech transmission over certain channels. Speech coders aim at reducing this bitrate, while introducing as little perceptual distortion as possible into the speech signal. Speech coders utilize the properties of human speech production and perception systems to achieve the bitrate reduction.

The six speech coders used in these experiments were chosen for their bitrate coverage of the coding standards and for their common use in telephone network applications. Tables 1 and 2 show the Signal to Noise Ratio (SNR) performance of these speech coders for a small sub-set of the TIMIT and ISOLET databases. The first column of the two tables with the title “1 coding” shows the SNR performance after coding a ‘clean’ signal through a speech coder once. The second column with the title “2 codings” shows the results after coding it a second time to show the effect of tandeming.

The SNR performance is calculated as follows:

$$SNR_{(dB)} = 10\ log_{10} \frac{\Sigma(S_{org})^2}{\Sigma(S_{org} - S_{coded})^2}$$

where S_{org} is the original (128 kbits/s) uncoded signal and S_{coded} is the signal that has been coded and decoded by a particular speech coder.

The first three coders are based on the backward adaptive differential pulse code modulating technique. These coders use linear prediction to remove the redundancy from the speech signal and are generally used for bitrates above 16 kbits/s. As the performance of this type of speech coder degrades quickly for bitrates less than 24 kbits/s, we chose three ADPCM coders above this bitrate for use in our experiments. Column 1 of Tables 1 and 2 lists the SNR performance of the ADPCM coders down to 24 kbits/s for a small set of the TIMIT and ISOLET databases. It shows that even at 24 kbits/s, the performance of the ADPCM coder is starting to decline rapidly. However, column

Table 1: SNR Performance versus bitrate of coders (1 Coding) and the SNR Performance by Tandeming (2 Codings) for the TIMIT database.

Type of	Bit Rate	SNR (dB)	
Coder	(kbits/s)	1 Coding	2 Codings
ADPCM G.723	40	25.89	24.37
ADPCM G.721	32	20.40	20.40
ADPCM G.723	24	16.19	16.19
LD-CELP G.728	16	15.88	11.55
GSM	13	12.41	10.39
CELP-1016	4.8	5.76	2.39

Table 2: SNR Performance versus bitrate of coders (1 Coding) and the SNR Performance by Tandeming (2 Codings) for the ISOLET database.

Type of	Bit Rate	SNR (dB)	
Coder	(kbits/s)	1 Coding	2 Codings
ADPCM G.723	40	21.41	20.10
ADPCM G.721	32	17.90	17.39
ADPCM G.723	24	13.80	14.00
LD-CELP G.728	16	14.39	10.44
GSM	13	7.67	6.83
CELP-1016	4.8	4.43	1.50

2 of Tables 1 and 2 shows that the SNR performance is affected very little by tandeming the ADPCM coder. This suggests that the recognition performance of the ADPCM coder should be very robust to tandeming.

The last three coders utilize the linear-prediction based anaylsis-synthesis procedure for coding speech. These coders model both the excitation source and production model and hence are more efficient for speech coding. The LDCELP and CELP-1016 are based on the CELP (Code Excitation Linear Prediction) method, whereas the GSM (Global System for Mobile Communications) coder uses the multi-pulse excitation model. The CELP coders are particularly useful for low bitrate environments.

In our experiments, we have used three commonly used CELP coders found in industry. The LDCELP coder descibed in [3] is a 16 kbits/s Low Delay CELP coder designed to perform better than the 24 kbits/s ADPCM coder and be comparable to the 32 kbits/s ADPCM coder. Tables 1 and 2 show that in terms of SNR performance, the LDCELP coder is comparable to or better than the 24 kbits/s ADPCM coder and slightly worse than the 32 kbits/s ADPCM coder. The LDCELP coder shows a relatively large degradation in SNR performance for both databases when used in tandem. This suggests that its recognition performance may be sensitive to tandeming.

The last two coders (GSM and CELP) are the lower bitrate coders. GSM is a standard developed for European digital mobile telephony applications, whereas the CELP coder is developed jointly by the U.S. Department of Defence and AT&T for secure transmission purposes. The SNR performance results for these coders are lower than the other coders as expected, due to their lower bitrates (see Tables 1 and 2). These coders also show a sensitivity to tandeming.

3. RECOGNITION EXPERIMENT

As stated in the previous section, the speech recognition performance is measured using two databases for the two recognition systems. The ISOLET database is used for the whole-word recognition system which consists of the 26 letters of the English alphabet. The training data consists of 4680 utterances from 90 speakers, and 1560 utterances from 30 speakers are used for testing. The database consists of the same number of female speakers as males and the speakers in the testing database are not used for training.

To test the phoneme based recognition system, we use all the training data from all the 8 dialects of the TIMIT database to train the 48 context-independent models as proposed in [4]. This database consists of 3696 training sentences from 462 speakers of which 326 are male. The testing is performed using the core test set suggested in the TIMIT documentation which includes 2 male speakers and 1 female speaker from each dialect resulting in 192 unique utterances from 24 speakers.

As both databases are sampled at 16 kHz with a 16 bit resolution, every utterance in the databases is decimated to 8 kHz using a low pass filter with a half power cuttoff of 3.5 kHz. The recognition systems are trained on the uncoded PCM speech data and tested using the coded speech utterances. The 128 kbits/s uncoded PCM data is used as a reference to show the performance degradation by using a speech coder at the input of a recognition system.

The HTK (HMM Tool Kit) package is used to observe the effect with both the whole-word and phoneme-based HMM recognition systems. Both systems use a simple left-to-right multi-mixture HMM model. Each of the whole word models contains 5 states, whereas the phoneme-based recogniser uses three states except for the closures and silence (epi, sil, vcl, cl) which are modelled with only a single state. All Guassian densities use diagonal covariances. The phoneme-based recogniser uses a bigram language model to improve overall performance.

In order to evaluate the effect of speech coding on speech recognition performance, we select two different feature extraction methods, LPC-derived cepstral coefficients (LPCEP) and Mel-spaced cepstral coefficients(MFCC). A 12th order LPCEP feature vector is initially compared to a feature vector of the same dimension using Mel-spaced cepstral coefficients. Later experiments use an enhanced feature set consisting of cepstrum, delta cepstrum and log energy.

In order to perform the feature analysis, the speech signal is analysed frame-wise (100 frames/sec) using a 20 ms Hamming window and pre-emphasis. Twelve LPC-derived cepstral coefficients are computed through a 14th order LPC analysis. Twelve Mel-spaced cepstral coefficients are obtained from 14 filter bank energies computed from the FFT-based power spectrum.

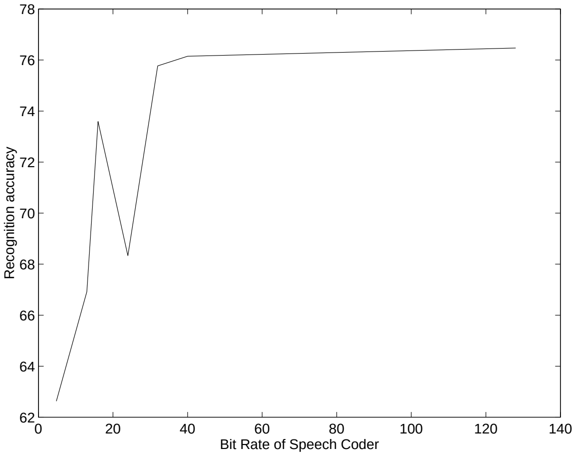

Figure 1: Typical recognition performance.

4. EXPERIMENTAL RESULTS

4.1. Effect of Coding

All the experiments conducted in this study show a decline in recognition accuracy with bitrate. This is shown in Figure 1 for a typical speech recognition system. Table 3 shows the results obtained for the ISOLET database using 12th order feature vector. As expected, the recognition performance decreases for lower bitrates.

Firstly, we notice that the MFCC coefficients are not affected as much by speech coding as the LPCEP coefficients. However, in both cases the LDCELP coder obtains a better result than the 24 kbits/s ADPCM G.723 coder and it is comparable to the 32 kbits/s ADPCM G.721 coder as reported in the literature [3]. These results are similar to the results obtained for the SNR performance using the same database (see Section 2).

Improved results are obtained by adding delta and log energy coefficients (see Table 4). Using 26th order MFCC coefficients, the performance degradation of the lowest bitrate coder (CELP-1016) compared to the reference is less than 6%. Without delta and log energy coefficients, this performance loss is nearly 10%. Similar results are shown for the LPCEP coefficients (see Tables 3 and 4). This shows that deltas and log energy coefficients are more robust to the distortion speech coders introduce into the signal. However, a small decline in recognition accuracy is still evident for decreasing bitrates.

To verify these results, we examined the effect speech coders have using a 26th order MFCC feature vector on a phoneme based recognition system. Table 5 shows a similar reduction in recognition performance with decreasing bit rate. However, as with the SNR performance in section 2 (see Table 2), the LDCELP coder did not perform better than the 24 kbits/s ADPCM coder using the TIMIT database.

Table 3: Recognition performance using the 12th order feature vectors and the ISOLET database.

Type of	Bit Rate	Recognition accuracy	
Coder	(kbits/s)	LPCEP	MFCC
No Coding	128	76.47	75.19
ADPCM G.723	40	76.15	75.32
ADPCM G.721	32	75.77	74.74
ADPCM G.723	24	68.33	72.88
LD-CELP G.728	16	73.59	74.74
GSM	13	66.92	73.27
CELP-1016	4.8	62.63	65.90

Table 4: Recognition performance using 26th order feature vector and the ISOLET database.

Type of	Bit Rate	Recognition accuracy	
Coder	(kbits/s)	LPCEP	MFCC
No Coding	128	88.14	87.63
ADPCM G.723	40	87.63	87.05
ADPCM G.721	32	87.44	87.44
ADPCM G.723	24	85.45	85.77
LD-CELP G.728	16	87.05	86.15
GSM	13	83.27	85.71
CELP-1016	4.8	80.19	81.86

4.2. Effect of Tandeming

In a typical communication system, the speech signal goes through a number of tandeming stages before it reaches the destination (e.g., the recognition system in the present study). Fig. 2 shows the processing of the speech signal with 2 stages of tandeming for a particular coder. In this paper, we investigate the effect of tandeming on recognition performance as a function of the number of tandeming stages. The training data for the recognition systems is the same as used in the previous experiments presented in this paper.

Recognition results as a function of the number of tandeming stages are shown in Tables 6 and 7. Here, we list the results for 2, 3, 4, and 5 tandeming stages. Results for 1 tandeming stage are already shown in Table 4 of the proceding section. It can be seen from these tables that tandeming of high bitrate coders does not effect the recognition performance much, while the recognition performance degrades significantly for the lower bitrate coders with tandeming. Degradation in recognition performance is more severe for the lowest bitrate coder. Also, as found in the experiments in the previous section, the LPCEP coefficients are affected more in the low bitrate environment and as shown in Tables 6 and 7, are also more sensitive to tandeming than MFCC.

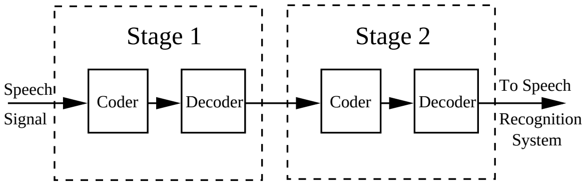

Figure 2: The processing of a speech signal with 2 stages of tandeming.

Table 5: Recognition performance using 26th order feature vector and the TIMIT database.

Type of Coder	Bit Rate (kbs)	MFCC
No Coding	128	52.63
ADPCM G.723	40	51.74
ADPCM G.721	32	50.84
ADPCM G.723	24	48.29
LD-CELP G.728	16	46.86
GSM	13	49.58
CELP-1016	4.8	45.23

Table 6: Effect of tandeming on recognition performance using a 26th order LPCEP and the ISOLET database.

Type of	Number of Tandems			
Coder	2	3	4	5
ADPCM 40 kb/s	87.56	87.56	87.50	87.50
ADPCM 32 kb/s	87.31	87.24	87.31	87.24
ADPCM 24 kb/s	85.19	85.19	85.19	85.19
LD-CELP	85.90	84.36	82.95	81.22
GSM	75.77	66.03	56.54	46.67
CELP-1016	66.28	55.77	44.17	36.60

As found in section 2 for the SNR performance, the recognition performance is not affected by tandeming the ADPCM coders. Also, the CELP-1016 and the GSM coders show a significant performance degradation as was found for the SNR performance (see Tables 1 and 2). The LDCELP coder however, shows only a small decrease in recognition performance even though the SNR performance showed it to be more sensitive to tandeming.

Results showing the effect of tandeming for the phoneme based recognition system on the TIMIT database are listed in Table 8. These results are similar to those reported earlier for the word recognition system. For example, the ADPCM coders show little degradation in performance. The GSM and CELP-1016 coders show a significant decrease while the LDCELP coder shows only a slight performance decrease.

5. CONCLUSION

As shown by their SNR performance, the ADPCM speech coders have little effect on speech recognition performance. The lower bit rate speech coders (GSM and CELP) have a significant effect on speech recognition due to the distortions they introduce. The LD-CELP coder showed better performance than the 24 kbits/s ADPCM coder and is comparable to the 32 kbits/s ADPCM coder.

Increasing the size of the feature set improves the overall performance. However, the recognition performance still degrades for smaller bitrate speech coders. The recognition performance reduced to 6%-8% below the reference compared to 10%-14% when using no delta or log energy coefficients.

Table 7: Effect of tandeming on recognition performance using a 26th order MFCC and the ISOLET database.

Type of	Number of Tandems			
Coder	2	3	4	5
ADPCM 40 kb/s	86.99	86.92	86.79	86.79
ADPCM 32 kb/s	87.44	87.37	87.37	87.37
ADPCM 24 kb/s	85.58	85.64	85.64	85.64
LD-CELP	86.28	86.09	84.23	82.50
GSM	83.14	79.74	75.32	70.26
CELP-1016	70.32	58.72	48.91	41.54

Table 8: Effect of tandeming on recognition performance using a 26th order MFCC and the TIMIT database.

Type of	Number of Tandems			
Coder	2	3	4	5
ADPCM 40 kb/s	51.57	50.84	50.55	49.94
ADPCM 32 kb/s	50.84	50.84	50.84	50.84
ADPCM 24 kb/s	48.29	48.29	48.29	48.29
LD-CELP	46.58	46.08	45.16	44.39
GSM	45.23	42.40	40.29	38.17
CELP-1016	39.38	34.97	29.36	27.07

The low bitrate speech coders (GSM and CELP) have a significant effect on recognition performance when tandeming. For 5 stages of tandeming, these speech coders degraded the performance by approximately 30%. The ADPCM speech coders virtually had no effect on the recognition performance while the LDCELP showed only a slight decrease.

Overall, the speech coders with a bitrate of 16 kbits/s and above displayed good recognition performance when using speech recognition systems with a 26th order feature representation.

6. REFERENCES

1. W.B. Kleijn and K.K. Paliwal (eds.), Speech Coding and Synthesis, Elsevier Science B.V., Amsterdam, 1995.
2. S. Euler and J. Zinke, “The influence of speech coding algorithms on automatic speech recognition,” Proc. ICASSP, Vol 1, pp. 621-624, 1994.
3. J.H. Chen and R. Cox, “The creation and evolution of 16 kbits/s LD-CELP: From concept to standard,” *Speech Communication*, pp. 103-111, June 1993.
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.