PLP COEFFICIENTS CAN BE QUANTIZED AT 400 BPS

Wira Gunawan and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, USA.

ABSTRACT

Previous work in wireless speech recognition has focused on two methods, namely, quantizing recognition features (e.g. MFCC) or performing recognition using speech coding parameters (e.g. LPC). All of this previous research assumes that the communication channel is only large enough to transmit either speech coding parameters or speech recognition parameters. By contrast, we propose that the speech recognition parameters can be quantized at a rate sufficiently low to allow transmission of both speech coding and speech recognition parameters over a standard cellular channel. In particular, this paper shows that the perceptual LPC (PLP) coefficients can be transmitted at 400 bps with an insignificant loss of digit recognition accuracy.

1. INTRODUCTION

Speech recognition is computationally expensive, requiring a large amount of memory and processing power. To overcome this problem, a speech recognition system may be distributed between the cell phone and base station. In the cellular phone, the front-end processor calculates, quantizes, and encodes speech recognition parameters. Encoded coefficients are transmitted to the base station, extracted by the decoder, and used in a speech recognition search.

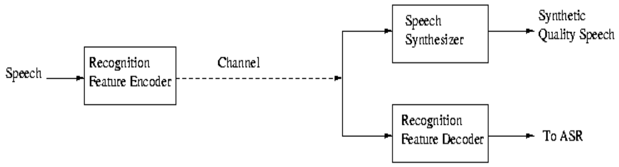

Figure 1: Class 1 distributed speech recognition: transmitter sends only speech recognition parameters. Speech for human listeners is synthesized from recognition parameters (after Kim and Cox [1]).

Kim and Cox discuss two different methods for distributing the computational complexity of speech recognition between the cell phone and base station [1]. In their first class of distributed system (figure 1), the cell phone transmits only speech recognition features to the base station, and speech for human listeners is synthesized from recognition features. This scheme yields recognition accuracy comparable to that of wireline ASR, as reported by Digalakis et al. [2]. Using MFCC (mel frequency cepstral coefficients) as a parametric representation of speech in the ATIS domain, they show that the required bit rate to achieve a recognition performance close to that of wireline ASR is 2000 bits per second. They reported a WER (word error rate) of 6.55% and 6.63% for wireline ASR and distributed ASR, respectively. The drawback of this scheme is that the synthesized speech does not have high quality because only speech recognition parameters are transmitted.

Class 2 systems, as shown in Figure 2, extract speech recognition parameters from the bit stream of the speech coder. Unlike the system in Figure 1, this scheme can produce wireline quality speech at the decoder side. Kim and Cox performed connected digit recognition experiments and reported a WER of 96.17% and 95.96% for wireline ASR and this approach, respectively.

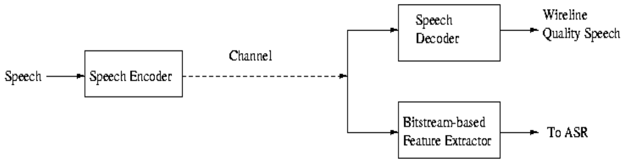

Figure 2: Class 2 distributed recognition: only speech coder parameters are transmitted. Speech recognition is performed using speech coder parameters (after Kim and Cox [2]).

In this paper, we propose a method, as shown in Figure 3, which quantizes both speech coding and speech recognition features, then mixes them for transmission and unmixes them at the decoder side. Transmitting recognition features as side information is feasible in our scheme because the bit rate of recognition parameters required to achieve high performance is very low. We obtain high recognition accuracy with only 400 bps using PLP analysis.

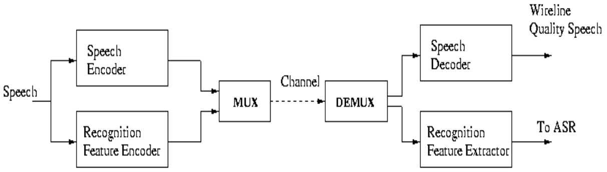

Figure 3: Both speech coding features and speech recognition parameters are transmitted in a single bit stream.

2. ALGORITHMS OVERVIEW

Perceptual linear predictive analysis (PLP) was proposed by Hynek Hermansky in 1989 [3]. PLP analysis is similar to linear predictive coding (LPC), except that the PLP technique also uses three concepts from the psychophysics of hearing. These three concepts are the critical-band spectral resolution, equal-loudness curve, and intensity-loudness power law (figure 4).

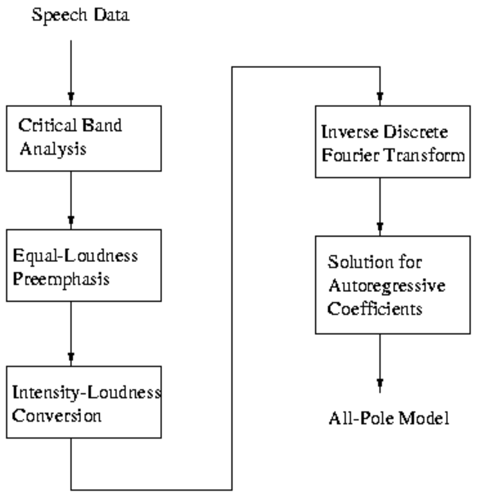
Speech Data
Critical Band Analysis
Equal-Loudness Preemphasis
Intensity-Loudness Conversion
Inverse Discrete Fourier Transform
Solution for Autoregressive Coefficients
All-Pole Model

Figure 4: Block diagram of PLP analysis (after Hermansky [3]).

Both LPC and PLP use the autoregressive all-pole model to estimate the short-term power spectrum of speech. However, as pointed out by Hermansky, the LPC all-pole model is not consistent with human auditory perception because it does not consider the nonuniform frequency resolution and intensity resolution of hearing. PLP alleviates this problem by applying the all-pole model to the auditory spectrum. The auditory spectrum is designed to be an estimate of the mean rate of firing of auditory nerve fibers.

For good word recognition accuracy, five PLP coefficients per frame are necessary. LPC-based speech coders typically use 10 coefficients in order to allow accurate synthesis of the four spectral peaks which are normally present below 4 kHz in a vowel spectrum. There is considerable evidence, however, that humans are not sensitive to the frequencies of all four spectral peaks. The psychophysical research of Chistovich [4] and others suggests that neighboring spectral peaks are merged into a single perceived peak if their frequencies are within about 3 bark of one another, and that vowel identification depends on the frequencies of only the first two perceived peaks. Low-order PLP analysis imitates Chistovich's psychophysical result: spectral peaks analysis with five coefficients can represent exactly two "perceived" peaks plus an overall spectral tilt. This theory is supported by Hermansky's experiments which show that the word recognition accuracy of a speaker-independent dynamic time warping (DTW) recognizer using PLP coefficients is best with five coefficients.

Dynamic time warping (DTW) is a nonlinear time-normalization algorithm for speech recognition based on dynamic programming [5]. DTW works by comparing a parametric representation of the input speech to stored templates. The stored templates contain the parametric representation of the vocabulary words. The parametric representation of speech used in this work is cepstral coefficients, which are derived from PLP coefficients. Pattern comparison is done by searching for the item in the templates that minimizes the Euclidean distance between the reference pattern and cepstral coefficients of the input.

3. EXPERIMENTS

Input speech waveforms are taken from the TIDIGITS database, and are downsampled to 8kHz prior to analysis. The first experiment is to do speaker-independent isolated digit recognition without any quantization. The test set consists of 220 utterances from 10 men and 10 women. The test set and the template set are completely different; in other words, no overlap between test files and reference files. Boundaries between silence and speech are calculated automatically and corrected manually. The spectral analysis uses a 30-ms asymmetrical window, which is the window used by the CS-ACELP speech coder [6]. The first 5/6 of the window is half of a Hamming window, and the last 1/6 is a quarter period of a cosine function. PLP analysis is performed every 10 ms. The order of the autoregressive PLP model is chosen to be five as discussed in the previous section. PLP coefficients are converted into cepstral coefficients before recognition is performed. In the first experiment, two templates per word are used to do recognition, i.e. the reference template consists of 22 utterances for 11 digits (two for each digit). Half of those are men's utterances, and the other half are women's. The accuracy of the speech recognition system in the first experiment is 93.18%.

The recognition accuracy is somewhat low for real applications and it is expected to decrease further as quantization is inserted into the system. Therefore, we increase the number of templates per word to 9 and 12. Hermansky did experiments by varying the number of templates per word from 2 to 23, and the recognition accuracy ranged from 92% to 98%.

Next, PLP coefficients are converted to LSP coefficients and quantized. The codebook size is varied from 64 (six bits) to 256 (eight bits) and the codebook is designed using the LBG algorithm. For all subsequent experiments, nine and twelve templates per word are used because the accuracy of the system with two templates per word is already too low and the quantization process will decrease the performance further. To further reduce the bit rate, PLP analysis is done every 20 ms instead of 10 ms, which is essentially down-sampling the speech pattern by a factor of two. Then the speech patterns are linearly interpolated to get the original sampling rate before recognition is performed. As LSP coefficients vary slowly from frame to frame, the distortion introduced by down-sampling and interpolation operations is small, so the degradation in recognition accuracy is not significant. Experiments using this scheme are also performed without quantization and with quantization using six to eight bits per PLP-LSP vector.

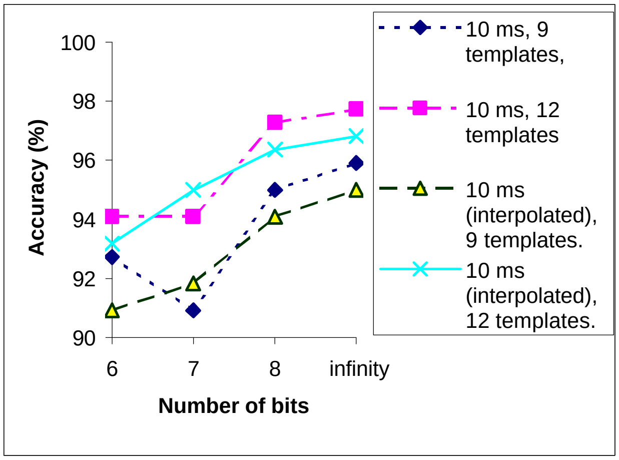

Figure 5: Recognition accuracy as a function of number of bits

Figure 5 shows the results of experiments discussed so far. PLP-LSP coefficients are quantized using 6, 7, or 8 bits per vector; an infinite number of bits corresponds to no quantization. As expected, the recognition accuracy declines as the number of bits used in quantization decreases, except for one case with a 10-ms analysis step. For all cases, 8-bit quantization yields only a slight degradation compared to no quantization.

Table 1: The recognition accuracy of experiments in this work.

Templates per word	Analysis step	Quantization			
		6	7	8	8
2	10 ms	-	-	-	93.18
9	10 ms	93.18	90.91	95.00	95.91
9	10 ms (interpolated)	91.82	91.82	94.09	95.00
9	20 ms	89.55	91.82	93.64	93.18
12	10 ms	94.55	94.55	97.27	97.73
12	10 ms (interpolated)	93.18	96.36	96.36	96.82
12	20 ms	91.36	93.64	94.09	95.00

Another scheme is to use 20-ms PLP analysis without linear interpolation to a 10ms rate. Both test and reference speech patterns have 20-ms analysis steps, so the number of computations is much less than that of previous experiments. Since much information is lost due to down-sampling, the recognition accuracy of this system is inferior to systems in previous experiments using the same number of templates per word. Without quantization, the accuracy only reaches 93.18% and 95% for 9 and 12 templates per word, respectively.

For convenience, all results of the experiments obtained in this work are tabulated in Table 1. To see whether the difference between algorithms using unquantized and quantized PLP coefficients is statistically significant or not, we perform McNemar's test [7]. Table 2 shows the significance level (*p*) of the performance difference between each quantized recognition scheme and the corresponding unquantized scheme. We can see from Table 2 that the difference between 8-bit quantization and no quantization is statistically negligible.

Table 2: Results of McNemar's test.

Templates per word	Analysis step	Quantization		
		6	7	8
9	10 ms	0.1435	0.0127	0.7539
9	10 ms (interpolated)	0.0118	0.1185	0.7539
12	10 ms	0.0215	0.0386	1.0
12	10 ms (interpolated)	0.574	0.424	1.0

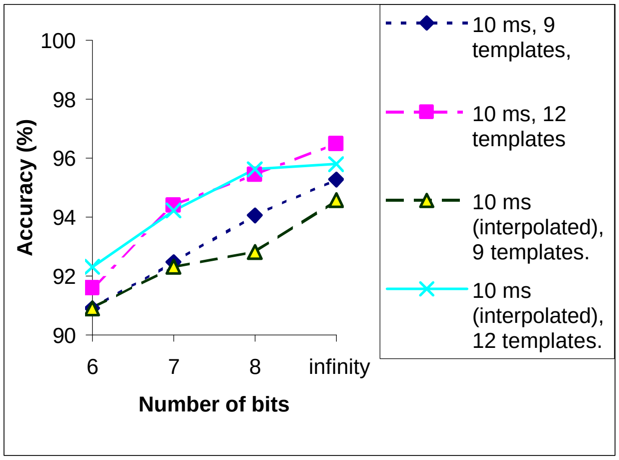

Figure 6: Recognition accuracy as a function of number of bits with larger database.

We also perform the experiments using a larger database with 572 utterances, with waveforms in which the speech endpoints are determined automatically with no manual correction. The automatic endpoint detection algorithm certainly is not as accurate as hand-edited endpoint detection, so we expect the recognition performance to be worse. Nevertheless, as we can see from Figure 6 and Table 3, the speech recognition performance with 12 templates per word does not degrade much. As the number of templates is decreased to 9 and 2, the decrease in recognition accuracy is larger; apparently the use of 12 templates per word compensates for some of the endpoint detection errors.

Table 3: The recognition accuracy of experiments using larger database.

Templates per word	Analysis step	Quantization			
		6	7	8	8
2	10 ms	-	-	-	89.34
9	10 ms	90.91	92.48	94.06	95.28
9	10 ms (interpolated)	90.91	92.31	92.83	94.58
9	20 ms	85.14	89.51	91.43	92.13
12	10 ms	91.61	94.41	95.45	96.50
12	10 ms (interpolated)	92.31	94.23	95.63	95.80

12	20 ms	86.89	90.73	92.48	93.36

We also apply McNemar's test to the new experiments with a larger database. Although the value of p decreases significantly for some cases, the 8-bit quantization scheme is still statistically similar to no quantization, as all p values for 8-bit quantization are above 0.05 (see Table 4).

Table 4: McNemar's test results using larger database.

Templates per word	Analysis step	Quantization		
		6	7	8
9	10 ms	0.00004	0.0025	0.1892
9	10 ms (interpolated)	0.0015	0.0708	0.0755
12	10 ms	0.0009	0.0755	0.2101
12	10 ms (interpolated)	0.0008	0.136	1.0

The results of the experiments suggest that a good speaker-independent digit recognition system can be developed using PLP analysis and the DTW algorithm with a fairly low bit rate. Of all experiments carried out in this work, the system employing down-sampling and interpolation seems most promising. If 8-bit quantization is used, the achieved bit rate is only 400 bit/s. This number is much lower than the bit rate of the speech coders used in wireless communication, which typically have a bit rate of 4.8 kb/s to 8 kb/s. Transmitting both CS-ACELP speech coder parameters and quantized PLP parameters would require 8.4 kb/s, an increase of only 5% over the normal CS-ACELP bit rate.

4. CONCLUSION

A speaker-independent isolated digit recognition system using PLP analysis and the DTW algorithm has been examined in this work. The system is examined with and without vector quantization. Based on the results of experiments, quantization using eight bits or higher is recommended since the degradation introduced in recognition performance is not significant compared to the system without quantization.

To minimize the bit rate further without degrading recognition performance, the analysis step at the front-end processor is changed to 20 ms. The experiment results show that the performance does not decrease substantially as long as linear interpolation is carried out at the decoder. A big advantage of this system is the very low bit rate required to transmit LSP coefficients; the bit rate is half that of a system with a 10-ms analysis step. Using 8-bit quantization for every 30-ms frame, the achieved bitrate is only 400 bps.

The proposed scheme might be applied by employing the G.723.1 standards [8]. 400 bps is less than the difference between the 5.3 kbps mode and 6.3 kbps mode in a G.723.1 multimode coder. A speech coder might therefore operate at 6.3kbps in its normal mode of operation. When simultaneous speech recognition is needed, the coder could change to a mode in which 5.3kbps are allocated to speech coder parameters, and 400bps are allocated to speech recognition parameters.

5. REFERENCES

[1] H.K. Kim and R.V. Cox, "Bitstream-based feature extraction for wireless speech recognition," in Proceedings ICASSP 2000, June 2000.

[2] V. Digalakis, L. Neumeyer, and M. Perakakis, "Product-code vector quantization of cepstral parameters for speech recognition over the www," in *Proceedings ICSLP*, 1998, pp. 2641-2644.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech, " *Journal of the Acoustical Society of America*, vol. 87 no. 4, pp. 1738-1752, 1990.

[4] L.A. Chistovich, "Central auditory processing of peripheral vowel spectra," *Journal of the Acoustical Society of America*, vol. 77 no. 3, pp. 789-805, 1985.

[5] H. Sakoe and S. Chiba, "Dynamic programming optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech, Signal Processing,* vol. ASSP-26, pp.43-49, February 1978.

[6] R. Salami et al., "Design and description of CS-ACELP: A toll quality 8 kb/s speech coder," *IEEE Transactions on Speech and Audio Processing,* vol. 6 no. 2, pp. 116-130, March 1998.

[7] L. Gillick and S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms" in *Proceedings ICASSP 1989*, vol. 1, May 1999, pp. 532-535.

[8] ITU-T, "Recommendation G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," March 1996.