PERSONAL SPEECH CODING

Wenhui Jia[1] *Wai-Yip Chan*

Department of Electrical and Computer Engineering
Illinois Institute of Technology
3301 South Dearborn, Chicago, IL 60616-3793
e-mail: {wjia chan}@ece.iit.edu

ABSTRACT

In existing speech coding systems, all quantizer codebooks are designed to suit the statistical and perceptual characteristics of speech signals of a population of speakers. However, an individual's speech signal does not exhibit, even over a long time, the entire range of characteristics of the population. With the advent of the personal communication systems, personal information might become available and be used to improve the rate-distortion performance of speech coders. In this paper we assess the potential gain of personal speech coding by designing codebooks for individual speakers. Spectral quantization, excitation and pitch lag codebooks of existing CELP coders are redesigned. The gains appear to be modest, suggesting that we need to use a different coding framework, which can model personal characteristics explicitly. Amongst the components, the spectral quantizer seems to be most amenable to personalization.

1. INTRODUCTION

Explosive growth in applications of digital wireless communications has provided a great impetus to research on low bit-rate speech coding. Although many advances have been made in speech coding technology, a significant leap in rate-distortion performance is still needed. In existing speech coding systems, all quantizer codebooks are designed to suit the statistical and perceptual characteristics of speech signals of a population of speakers. However, an individual's speech signal does not exhibit, even over a long time, the entire range of characteristics of the population. As the wireless communications network evolve into what we call the personal communications system (PCS), personal information might become available through the network or from a smartcard or a database. One way to improve the rate-distortion performance of a speech coder is to design the coder based on personal speech data. Therefore, an individual's speech could be encoded and decoded using his/her own codebooks. We call such a paradigm personal speech coding (PSC).

In the most general sense, a personal speech coder is a coder whose structure and codebook contents are completely tailored, by off-line design and/or online adaptation, to an individual's speech. As a first step, our work in this paper is limited to accessing the extent of rate-distortion efficiency enhancement that can be garnered from tailoring the contents of codebooks to individual speakers. The tailoring is done via off-line codebook design, and existing coding structures are used. In the rest of this paper, we first describe our experiment platform and speech database, and then we report our assessment of the personal gains of three coder components.

[1] Wenhui Jia is the recipient of a Motorola Partnerships in Research Grant.

2. PERSONALIZING LPAS CODERS

Linear prediction based analysis-by-synthesis (LPAS) coding is perhaps the most popular speech coding technique. In an LPAS coder, the characteristics of the speech signal are represented by three components: LP parameters, excitation signal, and pitch lag. We have performed experiments to assess the amount of personal coding gain, separately for each of the three components in an LPAS coding framework. Personal gain is expressed in terms of the amount of reduction in bit rate obtained by using personal codebooks instead of generic codebooks, while maintaining the coding quality furnished by the generic codebooks. Coding quality is assessed based on objective measures as well as informal subjective listening. The objective measures are average segmental signal-to-noise ratio (SegSNR, or simply SNR), average spectral distortion (SD), and average synthesized-speech spectral distortion (SSD). SD is widely used to measure the performance of LP parameter quantizers. SSD measures the net effect on the spectral envelope of reconstructed speech due to all speech coder components.

We set up a speech database consisting of datasets for individual speakers as well as for the general population. Personal speech datasets were constructed using speech recorded on commercial audio-book cassette tapes. Four speakers, two male and two female, were used. The speakers are given labels M-1, M-2, F-1, and F-2, where M stands for male and F for female. Each training set consists of about 55 minutes personal speech data, and each test set about 15 minutes or so. Another 8 speakers were chosen, each contributing about 8-minute of speech, to form our generic training set. There is no need for a generic test set in our PSC experiments. All speech material were sampled at 8 kHz, and low-pass filtered according to G.712 specification. We have also used the TIMIT database to represent a larger general population, but the recording condition is apparently different from the above 8-speaker generic set. For this reason, we do not assess personal gains using our

TIMIT-based results.

3. SPECTRAL QUANTIZATION

3.1. LP Parameter Quantization

For LP analysis, input speech is first highpass filtered and then processed using the lattice-LP analysis module of the GSM half-rate (VSELP) coder [2]. The module produces 10 reflection coefficients every 20 ms, using a rectangular window of 21.25 ms for analysis. The reflection coefficients are converted to direct-form filter coefficients, and thence to LSF (Line Spectrum Frequency) parameters. The LSFs are quantized and converted back to direct-form coefficients for use in speech synthesis. Two different "platforms" were used to synthesize speech for subjective evaluation. In the "LP-synthesis" platform, an all-pole synthesis filter using the quantized filter coefficients is excited by the original unquantized LP residual signal. Any audible distortion in the synthesized speech is due solely to spectral quantization. We also used the half-rate GSM VSELP coder as a synthesis platform. In this case, distortion in the synthesized speech is due to the combined effect of all coder components.

Table 1. SD performance results for LP quantization of personal test sets using personal codebooks

Test Data	Bits Used	Avg. SD (dB)	Outliers 2-4dB (%)	Outliers > 4dB (%)
M-1	27	0.897	0.65	0.01
	26	0.944	0.80	0.01
	25	1.010	1.51	0.01
	24	1.095	2.03	0.01
M-2	27	0.922	1.00	0
	26	0.968	1.13	0
	25	1.034	1.95	0
	24	1.118	2.75	0.01
F-1	27	0.872	0.48	0.01
	26	0.918	0.62	0.01
	25	0.975	1.02	0.01
	24	1.066	1.62	0.01
F-2	27	0.903	0.42	0.01
	26	0.950	0.57	0.01
	25	1.025	1.16	0.01
	24	1.098	1.76	0.00

After considering the trade-offs between performance and complexity, we elected to use three-split vector quantization (3-SVQ) to encode the LSF parameters. The SVQ groups the 10 LSFs in ascending order into subvectors of dimensions 3, 3, and 4. Separate SVQ codebooks were designed for the general population and for the individual speakers using the generalized Lloyd algorithm (GLA). The weighted Euclidean distance measure of Paliwal and Atal [1] is used. Their criterion for "transparent" quantization is also used: average SD of about 1 dB, less than 2% of 2-4 dB SD outliers, and no outlier having SD greater than 4 dB. Through experiment, we found that the following scheme for bit allocation is best in terms of SD performance. If the total number of bits of an SVQ is a multiple of three, then equal number of bits are assigned to the subvectors. Extra bits that can not be equally divided are allocated to the last two subvectors, first bit to the middle subvector, and the second to the last subvector.

Table 2. SD performance results for LP quantization of personal test sets using generic codebooks

Test Data	Bits Used	Avg. SD (dB)	Outliers 2-4dB (%)	Outliers > 4dB (%)
M-1	27	0.987	1.00	0.01
	26	1.037	1.18	0.01
	25	1.114	2.26	0.01
	24	1.204	3.27	0.01
M-2	27	1.055	2.72	0
	26	1.109	3.13	0
	25	1.181	4.46	0.01
	24	1.274	6.06	0.02
F-1	27	0.966	0.91	0.01
	26	1.017	1.08	0.01
	25	1.082	1.86	0.01
	24	1.185	3.05	0.02
F-2	27	1.035	0.90	0
	26	1.088	1.14	0
	25	1.168	2.34	0.01
	24	1.258	3.51	0.01

Tables 1 shows the quantization results obtained by applying the personal codebooks to their corresponding test sets. The SD performance versus bit rate is different across all the speakers. "Transparency" is attained at 25 bits for the two male speakers, and 24 bits for the two female speakers.

The above results are based on experiments wherein the training set and test set are both derived from the same source, e.g., codebooks trained on M-1 and applied to M-1. To estimate personal gains, we applied the generic codebooks to quantize the four personal test sets. The results are shown in Table 2. Clearly, SD performance is degraded in comparison with that shown in Table 1, which are based on applying the personal codebooks to quantize the personal test sets. Comparing the two tables, we see that for M-1 with generic codebooks, one or two more bits are needed in order to maintain nearly the same level of SD performance as using personal codebooks; thus, the personal gain is 1–2 bits for M-1. For all the other speakers, the gain is 2 bits. Based on Paliwal and Atal's criterion, "transparency" occurs at 26 bits for M-1 and F-1, 27 bits for F-2, and at more than 27 bits for M-2.

To get a sense of the degree of personal "tuning" of the codebooks, we "cross" the 25-bit personal SVQ codebooks with all the test sets to obtain the matrix of average SD results shown in Table 3. Every diagonal entry in Table 3 is the smallest amongst its corresponding row and column entries. Hence, the personal codebooks of each individual speaker are best matched to that speaker. Moreover, the codebooks of a different speaker but from the same gender produce a better match than across the gender.

Table 3. Average SD results for LP quantization by crossing 25-bit codebooks with all test sets (dB)

Codebook	Test Set			
	M-1	M-2	F-1	F-2
M-1	**1.010**	1.126	1.269	1.248
M-2	1.098	**1.034**	1.293	1.262
F-1	1.341	1.428	**0.975**	1.215
F-2	1.290	1.315	1.163	**1.025**
Generic	1.114	1.181	1.082	1.168
TIMIT	1.264	1.241	1.334	1.359

3.2. Listening Tests

To verify the objective gains, we carried out informal listening tests. Four sentences were selected for each speaker. For each sentence, we determined the bit rate at which "transparent" quantization occurs. "Transparent" quantization means that speech reconstructed using quantized LP parameters is indistinguishable from the original speech in listening. The largest bit rate used is 30 bits. To calculate an average bit rate at which "transparent" quantization occurs for each speaker, the highest and lowest rates out of the four rates for the four sentences are discarded and an average of the remaining two rates are taken. Based on this procedure, the number of bits required for "transparent" quantization using generic codebooks is 24, 25, 26.5, and 29.5 for M-1, M-2, F-1, and F-2, respectively. With personal codebooks, the numbers are 18.5, 20, 22.5 and 27 for the same order of speakers. The bit rate difference between the two cases for each speaker is regarded as the subjective personal gain: 5.5, 5, 4, and 2.5 bits for the four speakers. We note that the subjective personal gain for each speaker differs substantially from the objective personal gain. The objective results could be somewhat more reliable because the objective tests are based on much larger sample sizes.

3.3. LP Quantization on VSELP Platform

Besides the LP synthesis platform, we also assessed the subjective personal gain using the VSELP platform. We replaced VSELP's 28-bit FLAT quantizer with our 3-SVQ. First, we estimated the bit rate at which the generic SVQ is comparable to the 28-bit FLAT quantizer. Then, we did the same thing for the personal SVQs. By "comparable," we mean that the quality of reconstructed speech in the two cases is hardly distinguishable. The same set of four test sentences for each speaker and the same test scoring method is used, as in the subjective test using the LP synthesis platform. Our results show that the SVQ performs better than the FLAT quantizer. The improvement can be regarded as a "structural" gain, which is needed later on to calculate the personal gain. We found the "structural" gain to be consistent across the subjective listening results. First, the average gain for each speaker relative to the FLAT quantizer is determined. Then, we subtract off the structural gain for each speaker to arrive at the subjective personal gain over the VSELP platform: 3.5 bits for M-1, 3 bits for M-2, and 5.5 bits for F-1 and F-2.

4. EXCITATION CODEBOOK TRAINING

We used the FS1016 CELP coder to experiment with excitation codebook training. FS1016 has only one excitation codebook, containing 512 of 60-dimensional vectors [3]. The original code vectors were derived from a random number sequence. When we replaced the original excitation codebook with our own stochastic codebook populated from a Gaussian random sequence, we improved the reconstructed speech quality somewhat. The improved codebook served as the initial codebook for subsequent codebook training.

The basic excitation codebook training procedure follows the GLA framework: optimizing the encoder for a given decoder leads to partitioning or clustering of the training data, and optimizing the decoder for a given encoder leads to computing a centroid for each partition region. However, the excitation codebook is located inside the LPAS loop. The training is performed to minimize the average of the perceptually-weighted squared error (WMSE) of the reconstructed speech. Thus, training is based on speech data, not LP residual data. The resultant training procedure is "closed-loop." In this closed-loop procedure, the distortion measure contains weights that vary with the data, and the centroid computation is highly intensive [4, 5]. Though the speech data is fixed during training, clustering is actually performed on target data vectors. The target data vectors are derived from the speech data but they are also a function of the current excitation codebook. Consequently, the training set effectively varies from iteration to iteration in the closed-loop optimization. As a result, convergence is not guaranteed, even to a local optimum. In practice, however, the average WMSE does show large decrease in the first few iterations, as shown next by our results, and the training process can be stopped thereafter.

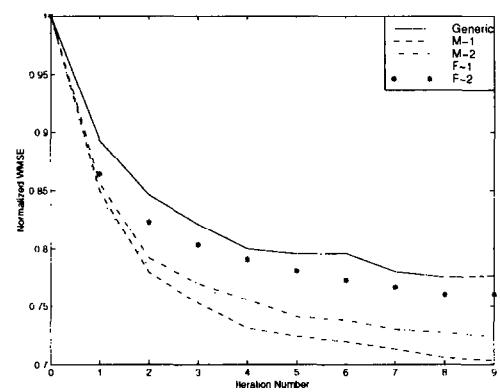

Figure 1. WMSE results for 9-bit excitation codebook training

During excitation codebook training, all other FS1016 codebooks are fixed except for the excitation gain, which we kept unquantized. With a subframe size of 7.5 ms in CELP, we chose to train over 8 minutes of speech so that the ratio of the size of the training set to the codebook size is 125. Similar to the LP quantization experiments, we designed a generic excitation codebook as well as four personal excitation codebooks. Distortion performance measurements obtained from the training process are plotted in Figs. 1-2. Training is terminated after ten iterations. For all personal

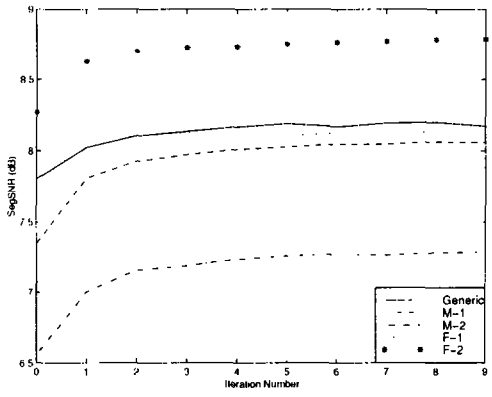

Figure 2. SNR results for 9-bit excitation codebook training

codebooks, the average WMSE decreases monotonically in the first eight iterations. The drop in WMSE is nearly 30% for M-1, and about 25% for the other speakers. For generic codebook, the drop in WMSE is not as much as in the personal cases. The average segmental SNR increases almost monotonically in all cases, providing an improvement of 0.3–0.7 dB.

Our listening tests showed that there is improvement in perceptual speech quality by using trained excitation codebooks, generic and personal, in comparison with using the stochastic codebook. However, the improvement is different across speakers. To estimate the personal gain in excitation training, we fixed the generic codebook at 9 bits, and reduced the number of bits for the personal codebook. Through listening, we determined at which point the reconstructed speech quality using the personal codebook is comparable to that using the generic codebook. The savings in bits due to using the personal codebook is a measure of the personal coding gain from excitation training. Personal codebooks of 6, 7, and 8 bits were designed. The final codebook for each case is selected according to the minimum WMSE criterion. The same sets of test sentences and the same scoring method as before were used to conduct the test. Listening tests showed a personal gain of 0.5, 1.5, 2 and 0 bit for M-1, M-2, F-1 and F-2, respectively.

5. PITCH LAG QUANTIZATION

In LPAS coding, the closed-loop pitch predictor serves to exploit long-term prediction gain rather than faithfully track the actual pitch periodicity of the speech signal. In VSELP, the allowable pitch lag range is from 21 to 142 samples, corresponding to a frequency range of 56 to 381 Hz, or almost three octaves. A typical speaker would produce about one octave of pitch range. To personalize the pitch quantizer, we set the pitch search range to a subrange of the original generic range. For example, by observing the histogram of pitch lag values used by VSELP in encoding personal speech, we found that the most heavily used portion of the pitch range for M-1 is 70-150 Hz, corresponding to lags between 53-110 samples. Then, only pitch values in this range are allowed in the encoding of M-1's speech, i.e., VSELP is allowed to search only in the pitch lag range of 53-110, rather than the full range. Thus, we could gain one bit by personalizing the pitch lag quantizer for M-1. Unfortunately, we found that the speech quality degrades drastically for most speakers. Hence, the pitch predictor provides long-term prediction gain that is critical to the operation of the coder. Another experiment we did that affirmed this role of the long-term predictor was to allow only male pitch range for female voice and vice versa. The gender identity was preserved in either case though quality was degraded.

6. CONCLUSION

We have reported on the experiments we have performed to evaluate the potential gain of personal speech coding. The results we have collected so far point to a modest rather than a drastic amount of personal coding gain. However, this tentative conclusion is arrived at without using any new coding structure or design algorithm, and for a very limited rate-distortion regime.

Our results thus far seem to suggest that we need to use a different coding framework, one that models personal characteristics explicitly. We give a very simple example to illustrate this point. If we were to design a PCM speech coder based on the statistics of the waveform amplitudes, we would probably arrived at coders of similar performance whether the design is based on statistics from a population or a person. Personal characteristic features might be salient at a higher level (e.g. prosodic) than that captured by the LPAS coders. Features extracted in "lower dimensions" may not be able to discriminate higher-level events. The fact that in our experiments we could substitute personal codebooks and pitch ranges from different speakers and still maintain much of the audible personal characteristics suggests that there is a great deal of redundancy in the LPAS coding platforms we have used, and not much personal characters were captured in the codebooks.

7. ACKNOWLEDGMENT

The authors would like to thank Dr. Michael McLaughlin and his speech group at Motorola for extending valuable help to this work.

REFERENCES

[1] K.P. Paliwal and B.S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Processing*, Vol.1, No.1, pp.3–14, January 1993.

[2] European Telecommunication Standard, *GSM 06.20 version 5.0.0*, December 1996.

[3] J.P. Campbell, T.E. Tremain and V.C. Welch, "The Federal Standard 1016 4800 bps CELP voice coder," *Digital Signal Processing*, 1, pp.145–155, 1991.

[4] G. Davidson, M. Yong and A. Gersho, "Real-time vector excitation coding of speech at 4800 bps," *Proc. ICASSP'87*, pp.2189–2192.

[5] J.-H. Chen, "High-quality 16 kb/s speech coding with a one-way delay less than 2 ms," *Proc. ICASSP'90*, pp.453–456.