SOURCE AND CHANNEL CODING FOR REMOTE SPEECH RECOGNITION OVER ERROR-PRONE CHANNELS

Alexis Bernard and Abeer Alwan

Dept. of Electrical Engineering, UCLA
Los Angeles, CA 90095
{abernard, alwan}@icsl.ucla.edu

ABSTRACT

This paper presents source and channel coding techniques for remote automatic speech recognition (ASR) systems. As a case study, Line Spectral Pairs (LSP) extracted from the 6th order all-pole Perceptual Linear Prediction (PLP) spectrum are transmitted and speech recognition features are then obtained. The LSPs, quantized using first-order predictive vector quantization (VQ) at 300 bps, provide recognition accuracy comparable to that of the baseline system with no quantization. A new soft decision channel decoding scheme appropriate for remote recognition is presented. The scheme outperforms commonly-used hard decision decoding in terms of error correction and error detection. The source and channel coding system operates at 500 bps and provides good digit recognition performance over a wide range of channel conditions.

1. INTRODUCTION

In this paper, we investigate source and channel coding techniques for remote speech recognition where the client extracts speech features and transmits them to the server for recognition. In most cases, such as in wireless transmission, the channel is error-prone. Previous studies have suggested alleviating the effect of channel errors by adapting HMM models [1] and ASR front-ends [2] to different channel conditions, or by modeling GSM noise and holes [3]. Other studies analyzed the effect of random and burst errors in the GSM bitstream for remote speech recognition applications [4].

We present here a novel channel coding technique specifically designed for remote ASR. The challenges in designing optimal source and channel coding techniques include keeping complexity low for the mobile client and minimizing the client-server transmission rate while providing high ASR accuracy for a wide range of channel conditions.

It is shown that speech recognition, as opposed to speech coding, can be more sensitive to channel errors than channel erasures. A novel channel coding technique optimized for error detection and including soft decision decoding of block codes is developed.

As a case study, Perceptual Linear Prediction (PLP) [5] ASR features are analyzed for source quantization. We design a source coder for the Line Spectral Pairs (LSPs) extracted from the PLP spectral representation. Isolated digit recognition experiments indicated that the source coder can operate at bitrates as low as 300 bits/s without degrading recognition performance.

Source and channel coding techniques are combined and we show good recognition results over a large range of channel conditions at overall bitrates as low as 500 bits/s. The soft decision

channel decoder, which introduces additional complexity only at the server, is proven to outperform the widely-used hard decision decoding for both error correction and error detection. The general framework presented can be extended to different ASR features.

2. SOURCE CODING CONSIDERATIONS

Feature vectors for ASR systems typically consist of spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) or Linear Prediction Cepstral Coefficients (LPCCs). LPCCs can be extracted from a standard linear prediction model or from a Perceptual Linear Prediction model (PLP) [5]. PLP systems model three properties of human auditory perception: critical band resolution, equal loudness, and intensity-loudness power law to derive an estimate of the auditory spectrum. PLP spectra can be represented using a low order all-pole model to yield a low dimensional representation of speech that is computationally efficient.

In the remainder of this section, we will analyze how to best quantize a 6th order all-pole representation of the PLP spectrum. The transfer function of PLP_6 is $A(z) = 1 - \sum_{k=1}^{6} \alpha_k z^{-k}$. One can obtain the LPCC (c_n) or the LSP (lsp_n) representation using:

$$c_n = \alpha_n + \sum_{k=1}^{n-1}(\tfrac{k}{n})c_k \alpha_{n-k} \qquad (1 \leq n \leq 6) \quad (1)$$
$$lsp_n = \text{roots}\{A(z) \pm z^{-(p+1)}A(z^{-1})\} \,(1 \leq n \leq 6). \quad (2)$$

The first question to address is whether LSPs or LPCCs should be transmitted and quantized. Quantizing LPCCs guarantees minimizing the Euclidean distance between quantized and unquantized LPCCs, assuring a close match between coded and uncoded feature vectors. On the other hand, LSPs typically improve coding efficiency, vary smoothly in time and hence can be linearly interpolated between sampled LSP values. This allows the LSPs to be updated more often than they are quantized.

In order to determine which information should be transmitted, we analyze three properties for both LPCCs and LSPs: 1) inter-frame correlation, which can be exploited with predictive coding to reduce the dynamic range of the information to quantize; 2) intra-frame correlation of the feature vectors, which results in coding efficiency when vector quantized; 3) sensitivity to quantization noise, which determine how good the quantizers should be.

The speaker dependent TI-46 digit database is used to compute correlations between LSPs and LPCCs of neighboring frames (20 ms apart). Results are shown in Table 1. Note the large and similar correlations for LPCCs and LSPs. An encoder can exploit this time redundancy by transmitting only the residual error after prediction.

Table 2 indicates the intra-frame correlations of the residual LPCCs and LSPs after prediction. Both auto-correlations are represented in a single matrix. The upper triangular matrix repre-

No.	1	2	3	4	5	6
LSP	0.87	0.88	0.88	0.91	0.92	0.90
LPCC	0.88	0.93	0.91	0.91	0.86	0.86

Table 1. Average (across all digits) inter-frame correlations between the six LSPs and LPCCs extracted from PLP_6 of adjacent frames, using 25 ms Hamming windows shifted every 20 ms.
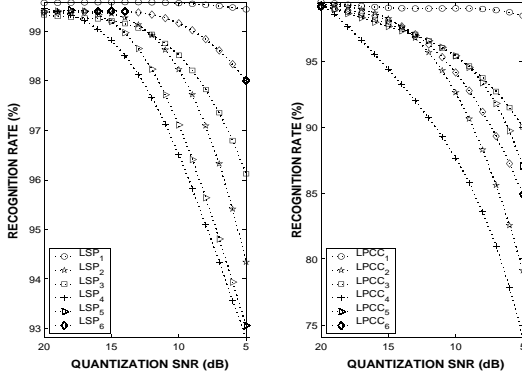

Fig. 1. Quantization error sensitivity analysis for the LSPs and LPCCs extracted from PLP_6

sents correlation of LPCCs and the lower triangular matrix, the intra-frame correlations of the LSPs. Large intra-correlations are observed for the LSPs, which can be efficiently exploited using vector quantization (VQ).

Fig. 1 illustrates the sensitivity of digit recognition results with respect to quantization errors when coding the LSP and LPCC residual after first-order prediction. Recognition was done on the TI-46 digit database (1180 male and female tokens for training, 480 for testing) and using HTK 2.1 with 5 states and 3 mixtures per word model. Note that the LPCCs are significantly more sensitive to channel errors than LSPs. Sensitivities also vary with the order of the LSP/LPCC feature. The individual sensitivities for each LSP will be taken into account in the training and search of the vector quantizer.

Based on these considerations, source and channel coding of the LSPs obtained from a 6th order PLP front-end is pursued hereafter given its low-dimensionality, low quantization error sensitivity and high inter and intra-frame correlation.

The six LSPs of PLP_6 can be quantized efficiently as follows: 1) remove the mean (DC component); 2) compute the residual LSP after a first order moving average prediction whose coefficient is

	Correlation of LPCCs				
1.00	-0.70	-0.11	-0.15	-0.16	-0.44
1.00	1.00	-0.19	-0.48	-0.03	-0.29
0.79	1.00	1.00	-0.15	-0.37	0.26
0.60	0.84	1.00	1.00	-0.10	-0.01
0.27	0.35	0.58	1.00	1.00	-0.04
0.05	0.11	0.16	0.63	1.00	1.00
0.17	0.20	0.29	0.38	0.69	1.00
	Correlation of LSPs				

Table 2. Intra-frame correlation of the residual LSPs and LPCCs after first-order prediction.
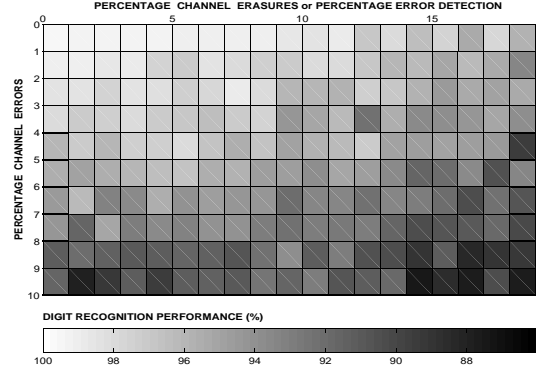

Fig. 2. Channel erasures and channel errors sensitivity analysis for the LSPs extracted from PLP_6.

chosen to minimize the signal variance after prediction; 3) vector quantize the residual vector using different one stage vector quantizers operating at 3, 4, 5 and 6 bits depending on the channel condition; the search cost function to be minimized is weighted depending on the error sensitivity of each LSP. The LSPs are transmitted every 20 ms and interpolated every 10 ms. This results in a total bitrate of only 150 to 300 bits/second. Table 3 reports digit recognition results at different bitrates.

3. CHANNEL CODING CONSIDERATIONS

The emphasis in remote ASR is recognition accuracy and not play back. Recognition is made by accumulating feature vectors over time and by selecting the element in the dictionary that is most likely to have produced that sequence of observations. The nature of this task implies different criteria for designing channel encoders than those used in speech coding applications.

For speech coding, frequent frame erasures due to poor channel conditions result in interruptions, buzzing and muting. For speech recognition where we accumulate observations over time, the situation can be different. Frame erasures reduce the number of observation vectors for all models, which may have little effect on recognition performance. Channel decoding errors, however, result in incorrect observation estimates, which in turn affect all state metrics accumulated in the Viterbi recognizer and can degrade significantly recognition performance.

Fig. 2 illustrates the effect of channel erasures and channel errors on digit recognition accuracy and confirms that recognition suffers more from channel errors than channel erasures. For example, it shows that if one can design a channel coder that limits channel errors to 1% or less and channel erasures to 10% or less, very high recognition accuracy can be obtained.

3.1. Linear block codes

In the previous section, we determined that the channel encoder protecting ASR features should provide reliable error correction

bits/frame	3	4	5	6
bits/sec.	150	200	250	300
Recognition	81.07	97.15	98.33	99.38

Table 3. Recognition accuracy after quantizing the LSPs of PLP_6 using mean removed first order predictive weighted VQ.

and error detection. Since the number of source information bits necessary to encode the LSPs of PLP_6 is low (K=4-7 bits per frame), block codes are favored over convolutional or trellis codes for delay and complexity considerations.

In order to guarantee the best possible recognition rate over a wide range of channel conditions, different block codes with different correction and detection capabilities are used. More source coding information bits will be used for high SNR channels while more bits will be used for channel coding in the case of low SNR channels. With such an adaptive scheme, graceful degradation in recognition performance is provided with decreasing channel quality. In the proposed design, Single Error Detection (SED), Double Error Detection (DED) or combined Single Error Correction/Double Error Detection (SEC/DED) codes are used depending on the channel conditions.

SED codes can be obtained using Cyclic Redundancy Check (CRC) codes. For instance, when a simple parity bit is added to the information codeword ($N = K + 1$), the minimal Hamming distance between valid codewords is $d_{min} = 2$. Any code with $d_{min} = 2$ is a SED code. We use two SED codes with parameters (8,7) and (8,6) to protect 7 and 6 information bits using 1 and 2 parity bits, respectively, for a total of 8 bits/frame or 400 bps.

However, when there are 2 errors among the N received bits, SED codes fail to detect the error and erroneous decoding is performed. To increase channel protection, we increase the number of source and channel bits transmitted to 10 bits/frame (500 bits/s) for intermediate and poor channel conditions.

For intermediate channel conditions, we use the $(10, 6)$ code generated by the polynomial $g(D) = D^4 + D^3 + D^2 + D + 1$. This code has $d_{min} = 3$ and guarantees that the smallest Hamming distance between two valid codewords is 3. With such code, one can decide to correct all single error events or to detect all one and two bits error events. Based on the above considerations we use it as a Double Error Detection (DED) code.

With decreasing channel quality, the number of errors detected increase rapidly, degrading recognition performance and it is necessary to correct and detect errors. A $(10, 5)$ SEC/DED code is obtained by expurgating the odd-weight codeword from a $(15, 11)$ Hamming code ($d_{min} = 3$) to form a $(15, 10)$ with $d_{min} = 4$. This code is then shortened to give a $(10, 5)$ code. The $(10, 4)$ code with $d_{min} = 4$ is obtained using the generator polynomial $g(D) = D^5 + D^4 + D + 1$. The dimension (N, K), minimal distance (d_{min}), and partial spectrum weights (A_i) of the linear block codes used are summarized in Table 4.

3.2. Hard vs. soft decoding

For a discrete memoryless channel, the probability of receiving the vector y given that the codeword x_m was transmitted is given by

$$p(y|x_m) = \prod_{j=1}^{N} p(y_j|x_{mj}) \quad (0 \leq m \leq 2^K - 1) \quad (3)$$

(N,K)	d_{min}	TYPE	A_0	A_1	A_2	A_3	A_4
(8,7)	2	SED	1	0	28	0	70
(8,6)	2	SED	1	0	12	0	38
(10,6)	3	DED	1	0	0	9	16
(10,5)	4	SEC/DED	1	0	0	0	16
(10,4)	4	SEC/DED	1	0	0	0	10

Table 4. Properties of the linear block codes used for channel coding the ASR features.

A decoder maximizing Eq. 3 without regard to the message *a-priori* probabilities is called a maximum likelihood decoder. This decoding rule is applicable to all discrete memoryless channel, including both hard- and soft-decision channels.

With hard decision decoding, the channel followed by the hard decision threshold acts like a binary symmetric channel (BSC) with cross probability $p = Q(\sqrt{2E_b/N_0})$, where E_b denotes the average energy per bit and N_0 the average noise energy. If channel noise statistics are stationary, the cross probability is a constant and the likelihood equation becomes $p(y|x_m) = p^{d_H}(1 - p)^{N-d_H}$, where d_H is the Hamming distance between y and x_m. Maximizing $p(y|x_m)$ is equivalent to minimizing the Hamming distance d_H between y and x_m.

For hard decision decoders, decoding is done as follows. For error correction, the decoding rule is minimum Hamming distance decoding, which achieve the minimum possible block decoding error probability. For error detection, the decoder detects an erasure when the Hamming distance between the received word and all valid codewords is non zero. For combined error correction and error detection (eg. SEC/DED codes), the decoding rule is to attempt error correction for small Hamming distances and to declare erasures for large Hamming distances.

Consider next a soft decision memoryless AWGN channel where the channel input is ± 1 and the channel output is a real number with Gaussian statistics. Specifically, the stationary channel is specified by

$$p(y|x_m) = \left(\frac{1}{\sqrt{\pi N_0}}\right)^N \exp\left[-\sum_{j=1}^{N} \frac{(y_j - x_{mj})^2}{N_0}\right] \quad (4)$$

Maximizing $p(y|x_m)$ is equivalent to minimizing the Euclidean distance $d_E = \sum_{j=1}^{N}(y_j - x_{mj})^2$ between y and x_m.

For soft decision decoder, decoding is done as follows. For error correction, The maximum likelihood decoder chooses its output to be the codeword x_m for which Euclidean distance between the received n-tuple and the codeword n-tuple is minimum. For error correction, soft decision decoding outperforms hard decision decoding. We propose here a new error detection rule for soft decision decoding of linear block codes that offers a continuous range of decision between attempting error correction and error detection. The technique works as follows.

For error correction, the codeword $\hat{x}$ exhibiting the smallest Euclidean distance d_{E_1} with the received vector y is selected. For error detection, we also compute the second smallest Euclidean distance d_{E_2} between y and the codewords x_m. If the relative difference in Euclidean distances is smaller than a threshold β (which indicates that two different codewords have a high probability of being sent), then the received vector y is not decoded and an error is detected. In other words, erasures are declared when

$$\frac{d_{E_2} - d_{E_1}}{d_{E_1}} < \beta. \quad (5)$$

Note that Eq. 5 is independent of the channel noise N_0.

As an example, consider an AWGN channel at 9 dB SNR. With hard decision decoding, the (10,5) SEC/DED code has a probability of Undetected Error (P_{UE}) of 1.7%, the probability of Error Detection (P_{ED}) is 16.1% and the probability of Correct Detection (P_{CD}) is 82.2%. These numbers are insufficient to provide good recognition results (Fig. 2). Fig. 3 illustrates the performance of soft decoding for the same (10,5) SEC/DED code operating at 9 dB SNR for different values of β. With soft decision decoding at $\beta = 0$ (which corresponds to error correction only), no
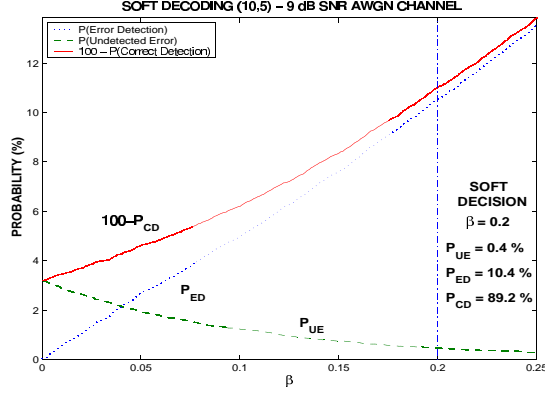
Fig. 3. Illustration of the soft decoding channel (10,5) SEC/DED channel decoder over an AWGN channel at 9 dB SNR.

error is detected ($P_{ED} = 0\%$) and $P_{UE} = 3.2\%$, which is larger than for hard decision decoding. With increasing β, however, one can rapidly reduce P_{UE} to the desired values (below that of hard decision decoding), while still keeping P_{CD} above that of hard decision decoding. For instance, with $\beta = 0.2$, $P_{UE} = 0.4\%$, $P_{ED} = 10.4\%$ and $P_{CD} = 89.2\%$, which can lead to good recognition accuracy. Note that when P_{UE} decreases, P_{CD} decreases as well, which indicates a tradeoff between the two probabilities.

The probabilities (correct decoding, undetected error and error detection) for the block codes designed for different AWGN channel SNRs are listed in Table 5. The value $\beta = 0.2$ is found appropriate to keep the number of undetected errors small while the probability of correct decoding remains high.

3.3. Recognition results with soft and hard decoding

Table 5 indicates that using soft decision decoding with $\beta = 0.2$ for both error correction and error detection applications leads to higher probabilities of correct decision and lower probabilities of undetected errors when compared to hard decoding. This in turn should lead to better recognition performance. Fig. 4 indicates the recognition performance improvement obtained by using soft decision decoding. Different block codes are used at different channel conditions and the overall transmission bitrate is 500 bps or less.

Note that soft decoding is made at the cost of additional complexity of computing Euclidean distances for all 2^K codewords. However, channel decoding is done at the server where the complexity of the recognizer prevails. The channel encoding operations for the client do not change.

(N,K)	SNR (dB)	P_{CD} (%)		P_{ED} (%)		P_{UE} (%)	
		Hard	Soft	Hard	Soft	Hard	Soft
(8,7)	13	89.9	94.0	9.6	5.6	0.5	0.4
(8,6)	12	82.1	91.1	17.4	8.5	0.6	0.5
(10,6)	11	94.4	95.8	2.2	3.9	3.5	0.3
(10,5)	10	89.2	96.1	10.0	3.6	0.8	0.3
	9	82.2	89.2	16.1	10.4	1.7	0.4
(10,4)	8	72.0	84.0	25.7	15.2	2.3	0.8

Table 5. Probability of correct detection (P_{CD}), error detection (P_{ED}) and undetected error (P_{UE}) using hard and soft decoding for the proposed codes used for different AWGN channel SNRs. Soft decoding is performed using Eq. 5 with $\beta = 0.2$.
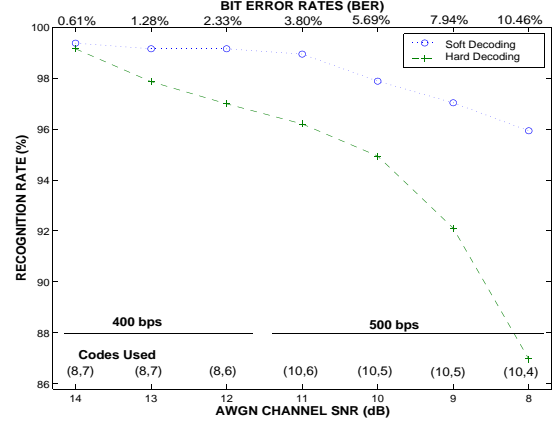

Fig. 4. Digit recognition performance using soft and hard decision over AWGN channel and BPSK modulation. Channel SNRs (dB) are shown in the bottom, estimated bit error rates on top.

4. SUMMARY AND CONCLUSIONS

In this paper, we present a framework for developing source and channel coding techniques for remote recognition systems. As a case study, we design a quantizer for the LSPs extracted from the 6th order all-pole PLP spectrum. Weighted predictive vector quantizer of the six LSPs at a rate of 300 bits/s provides recognition accuracy comparable to the unquantized system.

Simulations show that remote ASR is more sensitive to channel errors than channel erasures. Appropriate channel coding design criteria are determined. A set of low complexity linear block codes is shown to satisfy those criteria. A new soft decision channel decoding scheme that outperforms hard decision decoding is designed for both error correction and detection. The additional complexity is limited to the server. The overall source-channel coding system operates at 500 bps or less and provides good recognition performance over a wide range of channel conditions.

The source and channel coding techniques presented are not restricted to the transmission of LSPs and can be extended to other features such as MFCCs. Future work will examine other features as well as the effects of model size (word, phoneme, tri-phone) on quantization and channel protection design.

5. REFERENCES

[1] T. Salonidis and V. Digalakis, "Robust speech recognition for multiple topological scenarios of the GSM mobile phone system," in *Proc. of ICASSP*, 1998, vol. 1, pp. 101–4.

[2] S. Dufour, C. Glorion, and P. Lockwood, "Evaluation of the root-normalized front-end for speech recognition in wireless GSM network environments," in *Proc. of ICASSP*, 1996, vol. 1, pp. 77–80.

[3] L. Karray, A. Jelloum, and C. Mokbel, "Solutions for robust recognition over the GSM cellular network," in *Worksho Interact. Voice Techn. for Telecom. Applicat.*, 1998, pp. 166–170.

[4] A. Gallardo, F. Diaz, and F. Vavlerde, "Avoiding distortions due to speech coding and transmission errors in GSM ASR tasks," in *Proc. of ICASSP*, 1999, vol. 1, pp. 277–80.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoustical Soc. Amer.*, vol. 87, pp. 1738–52, 1990.