INITIAL EVALUATION OF MAXIMUM LIKELIHOOD CONTINUITY MAPPING FOR LOW BIT RATE ARTICULATORY CODING OF VQ CODE SEQUENCES

Patrick F. Valdez, John E. Hogden

CIC-3, MS B265
Los Alamos National Laboratory
Los Alamos, NM 87545
pfvaldez@lanl.gov, hogden@lanl.gov

Ramiro Jordan

Electrical & Computer Engineering
University of New Mexico
Albuquerque, NM 87131
rjordan@eece.unm.edu

ABSTRACT

A new approach to the compression of vector quantized (VQ) speech sequences is evaluated. The technique uses a method called maximum likelihood continuity mapping to learn a mapping between ariticulation and speech acoustics. Smooth articulator paths are then derived from VQ codes sequences. The paths are subsequently sampled, quantized, and transmitted along with additional information that allows perfect recovery of the VQ code sequences. A decoder takes the transmitted articulator paths and recovers the correct VQ code sequence for resynthesis of the speech waveform. The algorithm has not achieved compression yet, requiring an average of 6.04 bits/frame to transmit a 6 bit VQ code sequence, and 9.96 bits/frame to transmit a 10 bit VQ code sequence. However, modifications to the algorithm are currently under investigation, and we expect to implement improvements to help us compress VQ code sequences. Results of the improved algorithm will be presented at the conference.

1. INTRODUCTION

Many speech compression algorithms [1, 2, 3] transmit acoustic information through a sequence of vector quantization (VQ) codes [4]. For example, in the MELP system [3] line spectral frequencies (LSFs) [5] are derived from short-time (20–25ms) segments of the acoustic waveform, vector quantized, and transmitted as a sequence of VQ codes.

LSFs are based on the source-filter theory of speech, and convey information about the shape of the vocal-tract during speech production. Since speech sounds are produced by slowly moving articulators (tongue, jaw, lips, etc.), which can be transmitted using very few samples per second, it is reasonable to assume that articulator trajectories are better suited for coding than parameters of the acoustic waveform directly [6, 7].

Our goal is to take VQ code sequences and infer articulator paths for an encoder to transmit. The decoder should take the transmitted articulator paths and recover the correct VQ code sequences.

We use an algorithm called MAximum Likelihood COntinuity Mapping [8] (MALCOM) to estimate articulator paths from training data composed of VQ code sequences. MALCOM learns a *continuity map* (CM) that embodies a stochastic mapping between ariticulation and acoustics. MALCOM then finds articulator paths that maximize the probability of the observable VQ code sequences. Although the CM is found using only VQ code sequences, positions in the CM have been shown to be correlated with measured articulator positions [8, 9]. Thus, CM positions are called articulator positions throughout this paper.

This paper is organized as follows: MALCOM speech compression is covered in Section 2, the data is described in Section 3, the procedure applied to evaluate the algorithm is given in Section 4, followed by conclusions and future directions in Sections 5 and 6, respectively.

2. MALCOM SPEECH COMPRESSION

MALCOM assumes that the distribution of articulator positions, $\underline{x}'s$, used to produce VQ code c_j can be modeled by a Gaussian PDF, $p(\underline{x}|c_j, \underline{\Theta})$. Although articulator positions are unobservable, the techniques described below allow us to adjust means and covariance matrices of the Gaussians ($\underline{\Theta}$ comprises these parameters) to maximize the probability of the VQ code sequences. These techniques are similar in spirit to techniques used to optimize hidden Markov model parameters, in which the state sequences are also unobservable.

Since the "articulator" space is actually unobservable, sometimes we refer to it using a more neutral name: continuity map.

2.1. The MALCOM Model

In the MALCOM model, the probability of observing a code c_j given an n-dimensional CM position $\underline{x}$ is calculated with Bayes' rule as:

$$P(c_j|\underline{x}) = \frac{p(\underline{x}|c_j, \underline{\Theta})P(c_j)}{p(\underline{x})}. \quad (1)$$

where

$$p(\underline{x}) = \sum_i p(\underline{x}|c_i, \underline{\Theta})P(c_i). \quad (2)$$

Under the assumption of conditional independence, the probability of observing VQ code sequence $\underline{c}$ given an articulator path, $\underline{X}$, can be expressed as:

$$P(\underline{c}|\underline{X}, \underline{\Theta}) = \prod_{t=0}^{n} P(c_t|\underline{x}_t, \underline{\Theta}). \quad (3)$$

Note that Equation 3 does not make the claim that $\underline{x}_t$ is independent of $\underline{x}_{t-1}$; only that $\underline{x}_{t-1}$ gives no additional about c_t than already contained in x_t.

The MALCOM model also assumes that articulator trajectories are smooth; the trajectories have no energy above some cut-off frequency, f_c.

2.1.1. Inferring Articulator Paths

Ideally, the smooth path through a CM that maximizes the probability of the VQ code sequence is used as MALCOM's estimate of the articulator trajectory, i.e.

$$\hat{\underline{X}} = \underset{\underline{X}}{arg\ max}\ P(\underline{c}|\underline{X}, \underline{\Theta}) \quad (4)$$

However, because it is much faster to compute, we use

$$\hat{\underline{X}} = \underset{\underline{X}}{arg\ max}\ P(\underline{X}|\underline{c}, \underline{\Theta}) \quad (5)$$

2.1.2. Training Continuity Maps

Given the smooth paths that maximize the probability of the VQ code sequences, component parameters can be adjusted to increase the probability of the data.

$$\hat{\underline{\Theta}} = \underset{\underline{\Theta}}{arg\ max}\ P(\underline{c}|\hat{\underline{X}}, \underline{\Theta}) \quad (6)$$

A CM is trained by iterating between the following two steps: (1) given $\hat{\underline{\Theta}}$, find $\underline{X}$ to maximize $P(\underline{c}|\underline{X}, \underline{\Theta})$ and; given $\hat{\underline{X}}$, adjust $\underline{\Theta}$ to maximize $P(\underline{c}|\hat{\underline{X}}, \underline{\Theta})$. Note that a conjugate-gradient formulation is used to find $\hat{\underline{X}}$ and $\hat{\underline{\Theta}}$ of Equations 4 and 6.

In our implementation, we actually perform the following steps to train CM's: (1) given $\hat{\underline{\Theta}}$, find $\underline{X}$ to maximize $P(\underline{X}|\underline{c}, \underline{\Theta})$ then; given $\hat{\underline{X}}$, adjust $\underline{\Theta}$ to maximize $P(\hat{\underline{X}}|\underline{c}, \underline{\Theta})$ where the underlying component PDF's are modeled with symmetric Gaussians.

2.1.3. Compressing Articulator Paths

Since the articulator paths have no energy above f_c, they can be transmitted using $2f_c$ samples/sec. After uniformly quantizing the samples to b bit resolution, we use Huffman encoding to reduce the number of bits needed to transmit the samples.

2.1.4. Recovery of VQ Codes

The articulator path is not a lossless encoding of the VQ codes. However, given the articulator positions at time t, it is possible to determine the probability of each VQ code and rank-order the code probabilities. Thus, VQ code sequences can be transmitted losslessly by transmitting an articulator path along with the rank-order of the probability of the VQ code at time t given the path at time $t, \forall t$. For best compression, the rank orders are Huffman encoded. The bits used to transmit the rank order of the VQ codes are referred to as *error bits.*

3. DATA & SIGNAL PROCESSING

"Read" speech of one male speaker (British accent) from a book on tape was sampled at 8 kHz, digitized to 14 bits resolution, and used in our evaluation. The dataset consists of three stories; two stories were used as a training set, and the third story used as a test set.

Speech utterances were extracted from the digitized signal using an average adjusted magnitude (AAM) criterion [10], eliminating silence from further processing. After silence removal, the training set consisted of 175,996 20ms frames and the test set comprised 47,214 frames. This translates to approximately 58.7 minutes of training data and roughly 15.7 minutes of test data.

Utterances were initially bandpass filtered with an 8th-order Chebyshev Type II filter; passband frequencies 100 - 3200 Hz, stopband frequencies $\leq$ 50 Hz and $\geq$ 3600 Hz. We then applied a pre-emphasis filter ($\alpha = 0.95$), followed by a 10-th order LPC analysis (autocorrelation method) on nonoverlapping 20ms frames windowed with a 30ms Hamming window. Resulting LPC coefficients, a_k, were multiplied by $0.994^k, k = 1, ...10$ for a 15 Hz bandwidth expansion [11] of the formant frequencies. The bandwidth expanded coefficients, $a_k^{'}$, were then converted to LSFs.

4. PROCEDURE

4.1. Training

From the LSF encoded speech, 6, 8, and 10 bit VQ codebooks were constructed from the training data. The training data was subsequently VQ encoded producing VQ code sequences. These VQ code sequences were then used to find CM's with dimensions $D_i \in \{2D, 3D, 4D, 5D\}$, and cut-off frequencies, $f_c \in \{3Hz,$-$4Hz, 5Hz, 6Hz, 7Hz, 8Hz\}$. The number of MALCOM error bits needed to transmit the VQ code sequences in the training data were first calculated for unquantized articulator paths. Based on these preliminary results, we hypothesized that the 2D, 5Hz and 2D, 6Hz solutions of the 6 bit VQ encoded speech, and the 2D, 6Hz CM for the 10 bit VQ encoded speech gave us the best opportunity for compression. These CM's were analyzed further.

For each of these three CM's, articulator trajectories inferred from the training data were uniformly quantized using 2 bits/sample and a Huffman encoding of quantized articulator position samples was established. A second Huffman encoding was also found – an encoding of the rank-orders of the VQ code probabilities given the quantized articulator paths. Huffman encodings were also found using 3, 4, 5, 6, 8, and 10 bits/sample, for a total of 28 encodings. (2 encodings/quantization level, 7 quantization levels, 2 cut-off frequencies).

4.2. Results

VQ codebooks found on the training data were used to VQ encode the test data set. Smooth MALCOM paths were estimated using CM's derived from training data. Huffman encodings found on the training data were used to encode articulator paths and VQ code rank orders on the test data. The number of error bits required to losslessly recover the correct test VQ code sequences is given in Figure 1. Note that a change in quantization from 10 bits to 5 bits results in a modest increase in the MALCOM error bits; a coarser quantization of less than 5 bits, however, results in a significant increase of the MALCOM error bits.

Results of the average error bits/frame for the test set of 6 bit VQ code sequences are shown in Figure 2. With 6 bit VQ codes, MALCOM required an average of 6.04 bits/frame using the 2D, 6Hz CM and 3 bits/sample resolution for the path. A straightforward Huffman coding of the 6 bit VQ codes required 5.96 bits/frame. Likewise, MALCOM needed an average of 9.96 bits/frame to transmit the 10 bit VQ code sequences using a 2D, 6Hz CM and 4 bits/sample resolution of the ariticulator paths. Huffman coding of the 10 bit VQ codes directly required an average of 9.98 bits/frame.

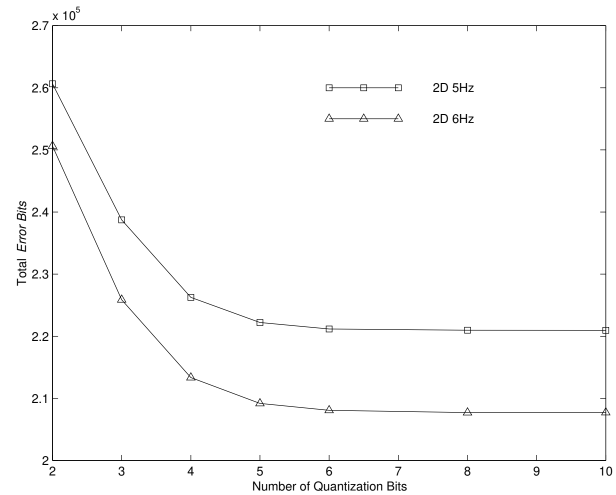

Figure 1: Quantization of Smooth Malcom Paths

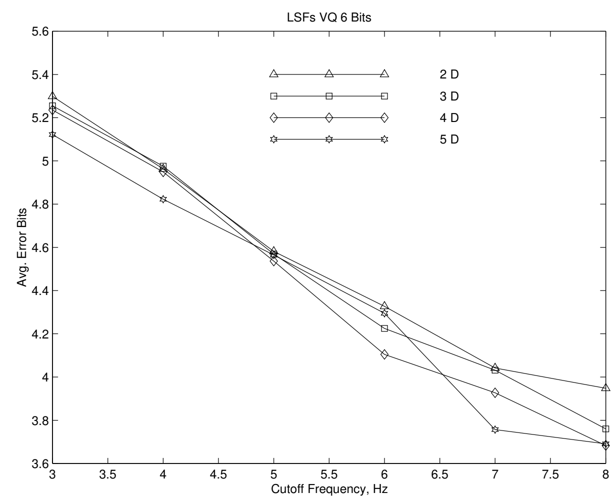

Figure 2: Average MALCOM Error Bits

5. CONCLUSIONS

We introduced a new approach to the compression of LPC parameters of the spectral envelope. This method uses articulator paths derived from LSFs to transmit the spectral envelope.

Our initial implementation of the compression algorithm has not achieved compression. At best, the algorithm approximately breaks even with a naive Huffman coding of the VQ codes.

The version of MALCOM used in this study models the CM as a mixture of symmetric Gaussians. It is conceivable, therefore, that with a more accurate modeling of articulator configurations in the CM, a resultant decrease in MALCOM error bits will give us compression. We are continuing to evaluate the algorithm on 10 bit VQ code sequences, and are exploring refinements outlined in Section 6.

It is also worth mentioning that since we have chosen the quantization level for the paths using test data, these results are likely to be somewhat better than results on a generalization test set. After further evaluation of the technique on the data described here, we expect to do a final test on new data.

6. FUTURE DIRECTIONS

We have identified several opportunities for improving the algorithm to achieve compression. Items under consideration include:

- Construct CMs with diagonal covariance matrices.
- Allow different dimensions of CM to have different cutoff frequencies; current implementations assign the same cutoff frequency to each dimension.
- Implement a more efficient quantization of smooth MALCOM paths. For instance, use a Lloyd-Max quantizer in lieu of uniform quantization.
- Consider a principal components representation of MALCOM paths in the frequency domain instead of bandlimiting in the time domain.
- Formalization of a procedure to help us choose the best CM for compression.

7. REFERENCES

[1] M. R. Schroeder and B. S. Atal. Code-excited linear prediction (celp): High-quality speech at very low bit rates. In *Proc. ICASSP-85*, pages 937–940, Tampa, FL, 1985.

[2] M. Nishiguchi, J. Matsumoto, R. Wakatsuki, and S. Ono. Vector quantized mbe with simplified v/uv division at 3.0kbps. In *Proc. ICASSP-93*, volume II, pages 151–154, Minneapolis, MN, Apr. 1993.

[3] A. McCree and et. al. A 2.4 kbit/s melp coder candidate for the new u. s. federal standard. In *Proc. ICASSP-86*, volume I, pages 200–203, Atlanta, GA, May 1996.

[4] J. Makhoul, S. Roucous, and H. Gish. Vector quantization in speech coding. *Proc. of IEEE*, 73:1551–1588, 1985.

[5] F. Itakura. Line spectrum representation of linear predictive coefficients. *JASA*, 54(4):S35, 1975.

[6] J. Schroeter and M. M. Sondhi. *Speech Coding Based on Physiological Models of Speech Production*, chapter 8, pages 231–267. Marcel Dekker, Inc., 1992.

[7] S. K. Gupta and J. Schroeter. Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis. *JASA*, 94(5):2517–2530, Nov. 1993.

[8] J. Hogden. Improving on hidden markov models: An articulatorily constrained, maximum likelihood approach to speech recognition and speech coding. Technical Report LA-UR-96-3945, LANL, Nov. 1996.

[9] D. A. Nix. *Machine-Learning Methods for Inferring Vocal-Tract Articulation From Speech Acoustics.* PhD thesis, Univ. of Colorado, Boulder, Jul. 1998.

[10] L. Rabiner and M. Sambur. An algorithm for determining the endpoints of isolated utterances. *Bell Syst. Tech. J.*, 54:297–315, Feb. 1975.

[11] R. Viswanathan and J. Makhoul. Quantization properties of transmission parameters in linear predictive systems. *IEEE Trans. on ASSP*, ASSP-23(3):309–321, Jun. 1975.