FILTER BANK BASED TRELLIS CODING OF SPEECH

Predrag Radivojac [1], *Vojin Šenk* [1], *and Milan Savić* [2]

1) University of Novi Sad, Department of Engineering, 21000 Novi Sad, Yugoslavia
2) University of Belgrade, Department of Electrical Engineering, 11000 Beograd, Yugoslavia
e-mail: radivojac@uns.ns.ac.yu; ram_senk@uns.ns.ac.yu; esavic@ubbg.etf.bg.ac.yu

ABSTRACT

A new procedure for trellis coding of speech is described. The algorithm uses a training sequence to find a long linear code for all stationary modes of speech and performs universal coding. Code search and encoding employ the M-algorithm that keeps a small fixed number of paths in contention. At rate 2 bits/sample we report a 4 dB improvement in signal-to-noise ratio over conventional trellis coding as well as over vector quantization.

1. INTRODUCTION

Trellis coding is a proven technique for source coding. It can be considered in terms of an encoder-decoder pair. The decoder consists of a finite state machine driving a table-lookup codebook of reproduction values. The encoder is a trellis search algorithm that chooses the channel sequence in such a way as to minimize the distortion between the input sequence and the decoder's output sequence. Generally, there is no structural restrictions on the finite-state machine and the most commonly used algorithm for the design of trellis source codes was proposed by Stewart *et al.* [1] which uses the generalized Lloyd algorithm and is based on a sufficiently large training sequence. Simulation results indicate that trellis encoders perform as well as or better than other coding systems for a relatively small number of states do. Anderson and Law [2] and Stewart *et al.* [1] reported that constraint length longer than 4 or 5 cannot improve the performance of the system. This results as a logical consequence of the fact that a communication system is to be designed for the speech source that is statistically ill-specified. This means that it has an inaccurate and incomplete probabilistic model.

Speech waveform is considered to be generated by a quasi-stationary source, i.e. its short-time behavior is stationary and ergodic but modes of stationarity occasionally change [3]. It is estimated that speech source has approximately fifty stationary modes with holding times ranging from 10 to 400 ms [2]. For such a class of sources, where, under some topological conditions, both on source and the distortion measure, the number of subsources can be approximated by a finite and relatively small collection of subsources, nearly optimal performance can be accomplished using universal coding. Universal coding refers to all techniques where the performance of the code selected without knowledge of the unknown true source converges to the optimum performance possible with the code specifically designed for the known true source. Technique of approximating a source with a finite set of stationary ergodic subsources and forming a union code is one of the universal coding techniques. The basic approach to universal code construction is then to select a finite number of representative subsources and to design a separate subcode for each. If the number of subcodes is small, then for a negligible increase in code rate a significant improvement can be obtained. The problem of universal coding is thoroughly analyzed in [4] [5]. Theoretically, such schemes are designed and are optimal for sources where nature chooses a subsource once forever, but it is also allowed to change subsources provided that they are switched slowly relative to the search time [4]. Unlike the common linear predictive coding, adaptation to each subsource means to

1) identify it in effect and
2) design an optimal corresponding trellis subcode.

All the subcodes that form the composite code are designed (or colored) prior to communication. When the communication starts, a path-map and a low-rate sequence, which specifies a code, are being sent. As the technology still rapidly goes further we consider a system with relatively large number of subsources at medium code rate of 2 bits/sample. Also, assuming that stationary modes can be distinguished and separated, we consider codes with large constraint length (memory) for which the rate-distortion limit, although not known exactly, can be reached. For such storage/constraint length requirements, a structural restriction only to real-number convolutional codes must be imposed.

2. REAL-NUMBER CONVOLUTIONAL CODES

It is well-known that the simplest class of convolutional codes for speech compression originates from the least mean-square (LMS) prediction, which is derived from the source autocorrelation function. One first calculates the long-term autocorrelation and the LMS prediction of a current source symbol and then incorporates them to a coding scheme as

$$v_t = \sum_{i=1}^{k} a_i v_{t-i} + x_t\,, \qquad (1)$$

where x_t represents the quantized difference of the source sample u_t and the first term on the right-hand side of (1); v_t represents the reproduced symbol at time t. More precisely, this expression is not a LMS prediction based coding scheme, it is a coding scheme inspired by LMS prediction. Prediction values **v** can also be obtained using the transversal filter $C(z) = 1/(1 - A(z))$, where $A(z)$ represents the Z-transform of the recursive filter with coefficients $a_1, a_2,\ldots, a_k$, computed from the first k lags of autocorrelation. It follows that

$$v_t = \sum_{i=0}^{\infty} c_i x_{t-i}\ . \qquad (2)$$

The memory length of **c** is potentially infinite, but the coefficient magnitudes in some cases fall off rapidly [2]. If $\mathbf{c} = c_0, c_1,\ldots, c_{\mathcal{M}}$ we refer to the code as a real-number convolutional code and we will employ them in the sequel. Codes based on LMS prediction are analytically obtained and are intended for single-path search procedures; in fact they are optimal for such procedures. For multi-path searches there are codes, obtained by computer search, that perform better.

Figure 1 represents a decoder for a real-number convolutional code. Coefficients of $C'(z)$ are normalized according to

$$C(z) = p \cdot C^{'}(z), \qquad |\,C^{'}(z)\,| = 1\,. \qquad (3)$$

However, codes based on LMS prediction have two basic shortcomings. First, all of them have infinite memory due to the recursive structure, and cannot simply be truncated; a prefix of a good long generator is not generally a good short generator (a possible improvement is truncation with tail memory). Furthermore, since LMS prediction, whose best approximation is DPCM, can be viewed only as a single-path search procedure, better performance can be accomplished using multi-path algorithms, with different codes. For such codes there is no analytical derivation and a computer search has to be carried out.

Although this is a standard optimization problem, Anderson and Law [2] devised a completely empirical procedure to find good codes for multi-path searches. This algorithm is based on an assumption that all stationary modes can be covered by a single short code. They reported that memory order longer than 5 cannot improve performance of a trellis encoder, and this is the result of averaging the autocorrelation function over all stationary modes so that the variation of the higher autocorrelation lags becomes too large and thus the variation of coefficients of $A(z)$ and $C(z)$. Nevertheless, such averaging is optimal for the design of a single code that covers all modes [6].

3. UNIVERSAL CODE CONSTRUCTION

To overcome this problem we first constructed an empirical procedure to extract stationary parts of a training sequence [7], i.e. to extract segments with approximately constant mean value and autocorrelation. This procedure gave us approximately wide sense stationary training segments. With this segments identified, we designed several good longer codes ($\mathcal{M} = 10$) for each training segment using optimization techniques such as the genetic algorithm, Anderson-Law's algorithm [2], and an empirical procedure proposed in [8] and M-algorithm as a trellis search procedure (with $M = 4$). Among these codes we extracted the set of representatives (subcodes) on condition that a speech segment cannot be encoded by a representative with loss greater than $\Delta = 1$ dB as compared to the code best matched to it. The steps of this algorithm are:

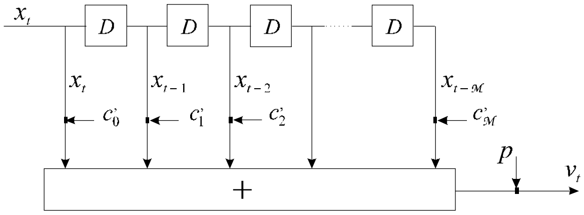

Fig. 1. The decoder of a real-number convolutional code

0) Let $\mathcal{S}$ be the set of wide sense stationary training segments, $\mathcal{C}$ the set consisting of several good codes for each segment from $\mathcal{S}$, and $\mathcal{B}$ an empty set containing codes accepted to be in a bank. Set $i = 1$.

1) Take i-th segment from $\mathcal{S}$ and find the best code from $\mathcal{C}$ matched to it. Let $SNR_{\mathcal{C},i}$ be the maximal SNR of the best code and $SNR_{\mathcal{B},i}$ the maximal SNR obtained using codes from $\mathcal{B}$. If $SNR_{\mathcal{C},i} > SNR_{\mathcal{B},i} + \Delta$, accept best code from $\mathcal{C}$ and store it in $\mathcal{B}$.

2) Set $i = i + 1$. If $i \leq card(\mathcal{C})$ go back to 1; else, stop the search. Set $\mathcal{B}$ is the set of representatives.

Unfortunately, such a procedure does not provide the minimum possible set of representatives in a single iteration. Assume we are given two segments $\mathcal{S} = \{s_1, s_2\}$ and corresponding best matched codes $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2\}$. If $\mathbf{c}_2$ represents segment s_1 within given distortion Δ of the reproduced sequence **v**, $\mathbf{c}_1$ may not represent s_2 within the same distortion. So, repeating the above procedure twice in the reverse order would certainly give the minimum set $\mathcal{B} = \{\mathbf{c}_2\}$. For large cardinality of $\mathcal{C}$ several iterations may be required.

The minimum set satisfying these conditions contains $B = card(\mathcal{B}) = 58$ codes, obtained in two iterations. Naturally, this filter bank can be extended to 64 in order to use binary transmission rationally. Encoding algorithm is very simple, although time consuming on a sequential machine. For a block of L source samples each of the B trellis encoders in a bank finds a scaling factor that provides the best matching of the input and reproduced sequence according to the mean-squared error. Having chosen the smallest mse, the best overall code is selected and the index of the best code, its scaling factor, and a path-map sequence are transmitted to the receiver.

Formally, the steps of this procedure are:

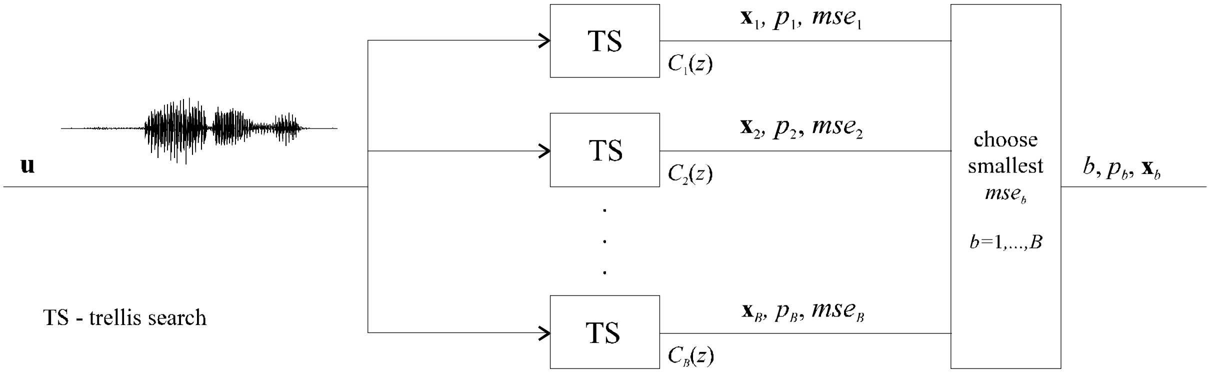

Fig. 2. Filter bank based source encoder

For each incoming speech segment containing L samples, repeat:

1) employing trellis search on a trellis generated by $C_i(z)$, find the mean-squared error mse_i, (or SNR_i) and its corresponding scaling factor p_i, $i = 1,\ldots, B$;

2) choose the smallest mse_i, and refer to it as mse_b (corresponding p is p_b and the path-map sequence $\mathbf{x}_b$), $b \in [1, B]$;

3) output the index b of the best code and the binary representations of p_b and $\mathbf{x}_b$.

This encoding system is shown in Fig. 2.

At rate R = 2 bits/sample (and B = 58 sets of coefficients in the bank) we have performed a simulation to compare full search memoryless vector quantization, conventional trellis coding, and our universal procedure. We have used the same training and test speech sequences 190,000 and 85,000 samples long, respectively. Standard VQ and trellis coding are simulated in order to compare various techniques on the same set of data.

Results are presented in Tables I-III and show the superiority of the new procedure for the negligibly increased code rate (6 bits per block as a side information - less than 3% of the total code rate). All results are in decibels.

Table I
Full search vector quantization, N - block length

N	SNR [dB]
1	7.16
2	10.75
3	12.23
4	12.79
5	13.23
6	13.41

In addition to signal-to-nose ratio (SNR), the performance of trellis coding is reported by tabulated segmental SNR ($SEGSNR$). For sampled speech, $SEGSNR$ is calculated by grouping the data into nonoverlaping segments, computing the SNR for each segment, and computing the arithmetic average over all segments with energy above some fixed threshold T. That is

$$SEGSNR = \frac{1}{K}\sum_{k=1}^{K} SNR_k \quad [\text{dB}], \tag{4}$$

where K is the number of segments with energy greater than T and SNR_k is the SNR of k-th such segment. Since the data contain no periods of silence, a threshold of T = 0 was chosen. Segmental SNR is considered to be a more realistic measure of speech quality.

Table II
Comparison of SNR-s for conventional and universal trellis coding; $\mathcal{M}$ = 10, M = 8.

L	Conventional trellis coding	Universal trellis coding
50	13.96	18.76
100	13.87	18.05
150	13.91	17.92
200	13.21	17.08
250	13.89	17.59
300	13.88	17.54
350	13.94	17.44
400	13.43	16.97

Table III
Comparison of $SEGSNR$-s for conventional and universal trellis coding; $\mathcal{M}$ = 10, M = 8.

L	Conventional trellis coding	Universal trellis coding
50	15.49	19.16
100	15.16	18.28
150	14.78	17.75
200	14.32	17.22
250	14.27	17.02
300	14.16	16.85
350	13.93	16.47
400	13.64	16.24

It can be seen from Tables I-III that signal-to-noise ratio for the conventional trellis coding remains practically the same while for universal coding it decreases with the block length. Eventually, for very large block lengths, these procedures would reach approximately the same performance. This performance would also reach memoryless vector quantization for sufficiently large block length N. The gain of the universal method is in the possibility to choose a more appropriate code for a given subsource. On the other hand, segmental signal-to-noise ratio falls off both for the conventional and universal coding. The reason for that is that the encoder searches for the path minimizing the mean-squared error. Consequently, the greater the block length the greater the adaptation to the higher energy segments. *SEGSNR* is the distortion measure that penalizes such adaptation. Nevertheless, the difference in performance due to universality remains. A slight decrease in performance of the universal code with the block length L arises from the fact that more than one of the stationary modes can be included in the same block, causing the overall decrease in both *SNR* and *SEGSNR*. The change of the parameter M in the M-algorithm to $M = 64$ adds approximately 0.5 dB to the *SNR* obtained with $M = 8$.

According to [9] we performed informal subjective MOS test. MOS scores require lengthy subjective testing, but are widely accepted as a norm for comparative rating of different systems. The rating scale employed in MOS testing is illustrated in table IV.

Table IV
Description in the Mean Opinion Score

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

Ten listeners were asked to classify perceived levels of distortion into either of the descriptive terms “excellent, good, fair, poor, unsatisfactory,” or into equivalent numerical ratings in the range 5-1. Comparing the original sequence with the reproduced ones obtained by classical and universal trellis coding the MOS score for the classically and universally coded sequence is 3.83 and 4.44, respectively. The reproduced sound is natural, easily understandable and with just perceptible distortion. That almost satisfies conditions for excellent communication quality.

4. SUMMARY

We have presented a procedure for designing universal trellis speech compression system based on filter banks. Major shortcomings of short nonadaptive trellis codes have been overcome by an adequately designed bank of trellis encoders based on stationary modes of the speech waveform. Results show an improvement in *SNR* of the order of 4 dB, for the same rate. Coding system shown here is raw in a sense that it just represents direct application of universal coding theory to speech source. We are sure that some kind of postfiltering and use of spectrally shaped distortion measure could improve subjective quality.

REFERENCES

[1] L.C. Stewart, R.M. Gray, Y. Linde, “The Design of Trellis Waveform Coders,” *IEEE Transactions on Communications*, vol. COM-30, NO. 4, April 1982.

[2] J.B. Anderson, C.W. Law, "Real-Number Convolutional Codes for Speech-Like Quasi-Stationary Sources," *IEEE Transactions on Information Theory*, vol. IT-23, pp. 778-782, Nov. 1977.

[3] J.S. Flanagan, *Speech Analysis, Synthesis, and Perception*, 2nd ed. New York: Springer-Verlag, 1972.

[4] R.M. Gray, “Time-Invariant Trellis Encoding of Ergodic Discrete-Time Sources with a Fidelity Criterion,” *IEEE Transactions on Information Theory*, vol. IT-23, NO. 1, Jan. 1977.

[5] A.J. Viterbi, J.K. Omura, *Principles of Digital Communication and Coding*, New York, McGraw-Hill, 1979.

[6] J.B. Anderson, J.B. Bodie, “Tree Encoding of Speech,” *IEEE Transactions on Information Theory*, vol. IT-21, pp. 379-387, July 1975.

[7] M. Savić, P. Radivojac, “A Heuristic Procedure for Extracting Stationary Modes of Speech,” *Proceedings of the XLI Yugoslav ETRAN Conference*, vol. 2, pp. 200-202, Zlatibor, 1997.

[8] P. Radivojac, S. Vučetić, “An Improved Trellis Code Search Procedure for Speech Compression,” *Proceedings of the XLI Yugoslav ETRAN Conference*, vol. 2, pp. 203-205, Zlatibor, 1997.

[9] S. Wang, A. Sekey, A. Gersho, “An Objective Measure for Predicting Subjective Quality of Speech Coders,” *IEEE Journal on Selected Areas in Communications*, vol. 10, NO. 5, June 1992.