Improving Recognition Accuracy on CVSD Speech under Mismatched Conditions

MADHAVI K. GANAPATHIRAJU, N. BALAKRISHNAN♦, RAJ REDDY
School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213
USA
balki@serc.iisc.ernet.in http://www.serc.iisc.ernet.in/personnel/balki.html

Abstract: Emerging technology in mobile communications is seeing increasingly high acceptance as a preferred choice for last-mile communication. There have been a wide range of techniques to achieve signal compression to suit to the smaller bandwidths available on mobile communication channels; but speech recognition methods have seen success mostly only in controlled speech environments. However, designing of speech recognition systems for mobile communications is crucial in order to provide voice enabled command and control and for applications like Mobile Voice Commerce. Continuously Variable Slope Delta (CVSD) modulation, a technique for low bitrate coding of speech, has been in use particularly in military wireless environments for over 30 years, and is now also adopted by BlueTooth. CVSD is particularly suitable for Internet and mobile environments due to its robustness against transmission errors, and simplicity of implementation and the absence of a need for synchronization. In this paper, we study some characteristics of the CVSD speech in the context of robust recognition of compressed speech, and present two methods of improving the recognition accuracy in Automatic Speech Recognition (ASR) systems. We study the characteristics of the features extracted for ASR and how they relate to the corresponding features computed from Pulse Coded Modulation (PCM) speech and apply this relation to correct the CVSD features to improve recognition accuracy. Secondly we show that the ASR done on bit-streams directly, gives a good recognition accuracy and when combined with our approach gives a better accuracy.

Key-words: CVSD, bitstream, speech recognition, corrected MFCC estimation

1. Introduction

A lot of effort has been put into signal compression on one side, in order to achieve transmission of speech at lower bandwidths, and into robust recognition of compressed speech on the other, in order to provide voice enabled command and control or natural language processing. While the former has seen significant amount of progress through the development of algorithms like CELP, RELP and CVSD that code speech intelligibly even at 2.4 kbps to 16 kbps, the performance of speech recognition systems has seen success mostly, only in controlled speech environments. Performance of speech recognition systems drops rapidly in noisy or compressed speech environments.

However, for the Automatic Speech recognition (ASR) systems to be of use, it is important that they perform well with speech signals that are used in practical transmissions. There have been efforts to study and improve the ASR performance on coded speech. Lilly and Paliwal have reported the effects of speech coders like ADPCM, CELP in [1] while Huerta and Stern have presented results on GSM coded speech in [2]. Besides these, Euler and Zinke also reported in [3], Sreenivasamurthy and Ortega in [4] and Zheng and Picone in [5], how the recognition accuracy drops in mismatched conditions between training and testing. Bit-stream based processing for ASR is reported in [6] and [7]. Most of the papers that discuss performance of ASR on coded speech study mainly LPC based coding schemes such as CELP and GSM.

♦ Corresponding Author.

In this paper, we present our studies of ASR performance on CVSD coded speech, which is becoming a preferred coding scheme for future mobile and handheld devices as explained below.

Coding algorithms such as CELP, MELP achieve good signal quality even at bit rates of the order of 2.4 to 13 kbps, whereas CVSD requires a larger bit rate, 16kbps or more, to achieve a similar quality. CELP, which is an algorithm of recent origin compared to CVSD, has been shown to be of better quality and more robust to background noise than CVSD [8]. While this is a drawback, CVSD has significant advantage that it does not require any special synchronization techniques, and the signal can be reconstructed starting from any intermediate sample from where signal is received. The other advantage of CVSD codec is its simplicity of implementation. It is simply a one-bit quantizer, with a simple adaptation, and hence can encode and decode speech in real-time. Table 1 gives comparisons on resources required for CVSD, LPC and CELP codecs, and it can be seen that the cost, memory, complexity and particularly, the delay, are 10 to 100 times less for CVSD. In robustness to transmission errors, it stands next to only CELP and has been in use in military wireless environments for quite sometime and is now being adopted for commercial wireless communications as well [9]. CVSD has also been accepted as the standard in the future wireless systems [10].

Unlike in Linear Prediction based coding schemes, CVSD bitstreams can be *directly* played to the speaker, without any decoding and reconstruction to get an intelligible quality speech. Motivated with this observation, we also performed speech recognition experiments on the bitstream treating it as a normal time domain signal, and got results that are comparable to those of decoded speech. This observation may be of particular interest to designers of Internet based application such as voice-chat and video conferencing, where robustness and minimal delay and lower complexity take high priority. It may also be used in server-based speech recognition systems, where any reduction to the already-high computational load would be useful.

The focus of this paper is to study the performance of automatic speech recognition on CVSD coded speech, and to evolve techniques that will offer improved performance for ASRs for CVSD coded speech signals.

	CVSD	LCP-10e	CELP(estimated)
Power	< 30 mW	500 mW	1~3 W
Cost	< $40	$650	$750-1500
Complexity	0.1 MIPS	1 MIPS	6.8-23.3 MIPS
Memory	1k ROM	6k ROM	8k ROM
Delay	0.25 ms	157.5 ms	120 ms

Table 1 Resource Requirements for 3 different coding algorithms, (Source: [8], only partially reproduced here)

2. CVSD Encoding and Decoding

CVSD is a form of Adaptive Delta Modulation (ADM), where a sample of the signal is coded as a 1 or 0. The characteristic feature of the CVSD among the various ADMs is that the step size δ is updated based on the previous 3 or 4 samples [11].

The simplest of Delta Modulations, Linear Delta Modulation (LDM), may be explained by the following equation:

$$(1) \quad b[n] = \mathrm{sgn}\,(x[n] - y[n-1])$$

where x[n] is the original sample, y[n-1] is the previous reconstructed sample and b[n] is the encoded bit. y[n] is reconstructed as follows:

$$(2) \quad y[n] = y[n-1] \pm \delta$$

depending on whether b[n] is 1 or 0 and δ is fixed step size.

Continuously Variable Slope Delta (CVSD) modulation is an adaptive version of delta modulation, where the step size δ is varied based on the past 3 or 4 bits. The basic algorithm for CVSD may be explained by the following equation, as given in [11]

$$(3) \quad \delta[n] = \beta\,\delta[n-1] + \alpha[n]\,\delta_0; \quad \beta = 1-\varepsilon^2 \text{ with } \varepsilon \to 0$$

where, $\alpha[n] = 1$ if the K bits out of the last J bits have been 1 or 0, and $\alpha[n] = 0$ otherwise. δ_0 is a fixed step size at which δ is incremented or decremented. This is a version of (J,K,L) algorithm, also described in [11]. δ is also constrained by δ_{max} and δ_{min}, as the maximum and minimum values that it may attain. δ_{max} is achieved automatically as (3), becomes a geometric

progression with multiplicative factor $\beta<1$ in the case of continuous decrease of δ, and this ensures that it is bounded above. The minimum value of δ, $\delta_{min,}$ is enforced over (3), and when this condition is reached, the algorithm degenerates to LDM, until the input signals an increase of the step size δ. δ_0 is usually same as δ_{min}. δ_{min} and δ_{max} are chosen depending on the dynamic range, the maximum frequency and sampling frequency of the input signal. β is chosen based on the syllabic time constant for step size decay, *i.e.*, the time taken for step size to decrease by a factor of 1/*e*. From (3), the syllabic time constant τ and β are related as

$$(4) \quad \tau = [\, f_s \ln (1/\beta)]^{-1}$$

Staircase integrator described in (2) is also typically modified to

$$(5) \quad y[n] = h * (y[n-1] \pm \delta)$$

where *h* is related to the integrator time constant t as

$$(6) \quad t = [\, f_s \ln (1/h)]^{-1}$$

Typical values for the syllabic and integrator time constants are 5 to 12 ms and 0.5 to 1.5 milliseconds, respectively. The quality of the decoded signal is very closely related to the parameters β, *h* and δ_0, which in turn are determined based on the input signal characteristics. The design of CVSD codec is described at length in [11].

3. Experimental Set up and baseline results

We designed the CVSD codec as described above and implemented it in software. All the experiments were conducted using the CMU Sphinx-III system on the TIDIGITS data set. The training set was made up of 8620 utterances and testing was performed on 8600 utterances. The original data consists of speech recorded in studio environment, sampled at 20kHz and stored with 16 bits per sample resolution. These speech utterances were down-sampled to 16kHz to make the *PCM* data set. The PCM was then converted to *BIT* which was the bit stream at 16 kbps obtained by CVSD encoding. BIT was then decoded by CVSD decoder, to yield the *CVSD* data set.

The results of the baseline experiments are shown in Table 2. The word error rate for CVSD tested with PCM models is as high as 29%. In the next section we present two novel techniques to reduce the word error rate and improve accuracy.

4. Approaches to improve recognition accuracy over CVSD speech

We propose the following approaches to improve the recognition accuracy for CVSD speech.

4.1. Add CVSD data to the training data

We tested the recognition accuracies with acoustic models trained with CVSD alone, and also with those trained with both CVSD and PCM data. The results are presented in Table 2. It can be seen that the accuracy for word and sentence recognition has improved from 92% to 98.4 and from 71.4% to 91.4% respectively. Though these improvements are significant, they are still unsatisfactory for any practical application.

Training Set	Testing Set	Bitrate (kbps) (Test set)	Recognition Accuracy %	
			Words	Sentences
PCM	PCM	256	99.2	94.8
PCM	CVSD	16	92	71.4
CVSD	CVSD	16	98.4	91.4

Table 2 Recognition accuracy for baseline and matched case.

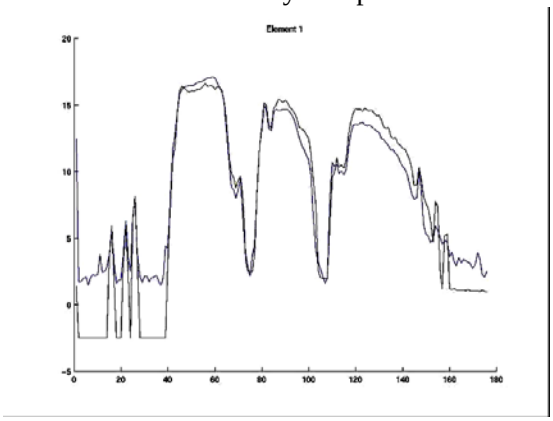
When the training set was expanded to include the PCM data as well, the performance for recognizing CVSD coded speech degraded a bit, but still offering performance levels that are better than the CVSD recognition using the HMM trained using the PCM samples alone.

4.2. Correct the MFCCs of CVSD by linear regression

In all the ASR experiments, we used 13 Mel Frequency Cepstral Coefficients (MFCCs) to generate feature vectors. We compared the MFCCs generated from PCM data and those generated from CVSD data. A few typical segments are presented in Figure 1 and 2. The observations that can be made are as follows:

The 1st of the MFCCs is almost identical in both cases. The 2nd, 3rd and 4th elements of the CVSD speech are

scaled down versions of the corresponding elements of PCM speech (only 1st and 4th MFCCs are shown in figures here). This can be attributed the fact that CVSD bits are chosen by comparison of the values of



the samples to the previous sample, thus being less sensitive to the exact scaling.

Figure 1 1st MFCC elements of PCM (blue) and CVSD (black)

The deviations of the rest of the elements in the vector are not significant. These belong to higher frequencies wherein the energy in the speech itself is low.

Figure 1 shows the lowest band MFCC (element 1) of CVSD and PCM computed for a sample file. It can be easily seen that there is not much difference between the two. Whereas, it can be seen in Figure 2, which shows a corresponding information for the 4th band (element 4), that the CVSD component is a scaled down version of the PCM component.

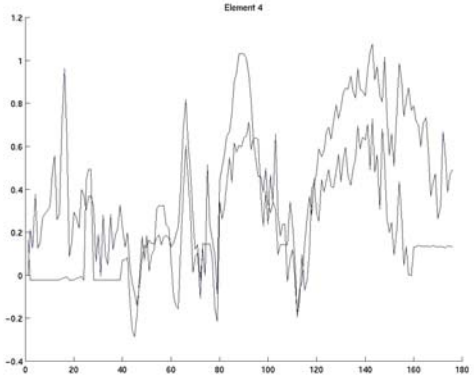


Figure 2 4th MFCC elements of PCM (blue) and CVSD (black)

Further analysis of all the MFCCs showed that the feature vectors of CVSD and PCM are linearly related. Given a set of MFCCs of the clean speech, derived using PCM samples, one could estimate the values of the MFCCs derived from CVSD speech as a linear combination of the elements of the former. Or, for a given, set of MFCCs derived from CVSD speech signal, those *would have been* for PCM may be *estimated*, as in:

$$P_2 = \alpha_{1,2}C_2 + \alpha_{1,3}C_3 + \ldots + \alpha_{1,13}C_{13}$$
$$P_3 = \alpha_{2,2}C_2 + \alpha_{2,3}C_3 + \ldots + \alpha_{2,13}C_{13}$$
$$\ldots \quad = \quad \ldots$$
$$P_{13} = \alpha_{13,2}C_2 + \alpha_{13,3}C_3 + \ldots + \alpha_{13,13}C_{13}$$

Table 3 Approximating MFCC element corresponding to PCM as a linear combination of the MFCC elements of CVSD

We used a sample of 50 files from the training set to compute the coefficients α's given in Equations 3-5 above. Using these, we computed the estimated PCM values (CVSD_E) from the CVSD training and test data. This correction in the MFCC features has yielded a recognition accuracy of 96.8% when tested with PCM trained acoustic models, which is over 50% reduction in word error rate. The recognition accuracy is given in Table 4.

Training Set	Testing Set	Bitrate (kbps) (Test set)	Recognition Accuracy %	
			Words	Sentences
PCM	CVSD_E	16	96.8	80.8

Table 4 Recognition Accuracy with the proposed modification of MFCC.

5. Results of Bitstream based feature extraction

CVSD is a sample based coding scheme, unlike CELP, RELP, etc, which are frame based coding schemes. Hence, the bitstream may not only be used to extract features required for ASR, but may also be used for direct play back. This makes it a potential candidate for speech applications in low-power devices and for applications that are likely to have transmission losses such as voice-chat and video conferencing in mobile and Internet applications.

With these interests, we quantified *how good* the quality of bitstream is, by performing ASR on features extracted from the use of CVSD bit streams directly. The results are presented in Table 5. The recognition rate using models trained with PCM samples was 63.2% and 40.1% respectively for word and sentence recognition when the bit sequences were used for testing. This is expected since the bit sequences have the non-linear effects of step size adaptation embedded in them and hence are not directly related to the MFCCs. The recognition accuracy improved to 90% (words) and 76% (sentence) when it was used with the models trained using the CVSD data. This indicated that the MFCCs derived using the CVSD decoded data is closer to the bit sequence. This prompted us to try the approach of using the CVSD bit stream directly for training and testing. This improved the recognition accuracy to 94.1% for words and 81.7% for sentences. This is comparable to what one could obtain using the CVSD decoded speech. In many simple hand held applications, wherein the complexity of the CVSD decoder may be too much to embed, the direct use of the bit streams may offer acceptable recognition accuracy with a much simpler hardware.

6. Performance in the presence of transmission errors

Transmissions errors are normally recovered by channel coding methods, which employ error-correcting codes. Hence, bit errors are usually not a problem for speech communication. We can comment that CVSD suffers only deletions in the presence of frame drops, since its adaptation is fast, and since encoding is done on a sample-by-sample basis.

We also studied how CVSD performs in presence of bit errors. We found that in the presence of irrecoverable bit-errors, decoded speech performs better than bit-stream based recognition. The results for decoded speech are summarized in Table 6. The percentage of errors introduced by selecting a uniformly distributed bit positions and flipping them to 0 to be "lost bits". It shows that for an x% increases in bit errors, the word error rate (WER) increases about 3x to 4x times.

Training Set	Testing Set	Bitrate (Test set)	Recognition Accuracy	
			Words	Sentences
PCM	BIT	16	63.2	40.1
CVSD	BIT	16	90	76
BIT	BIT	16	94.1	81.7

Table 5 Recognition Accuracy for features extracted from Bitstream based.

Training Set	% errors in Testing Set	Bitrate (Test set)	Recognition Accuracy	
			Words	Sentences
PCM	2% errors	16	84.529	65.4
PCM	5% errors	16	71.28	40.4
PCM	6% errors	16	58.9	40.0
CVSD	2% errors	16	94.86	84.3
CVSD	5% errors	16	86.66	60.3
CVSD	6% errors	16	80.0	59.0

Table 6 Recognition Results of CVSD decoded speech in the presence of bit errors

7. Conclusions and future work

We have shown that the recognition error rate doubles when speech undergoes an encoding and decoding as against that tested in matched conditions. Whereas, in mismatched conditions where the acoustic models are those of clean speech, we have shown that re-estimating the features obtained from coded speech as a regression on clean speech decreases the error rate by over 50%. The recognition accuracy achieved by this re-estimation is almost as good as that achieved in matched conditions against acoustic models created with encoded speech. Though this technique has been demonstrated for CVSD alone in this paper, similar approach may be applicable to other coding methods such as CELP, MELP, etc.

We have also presented the recognition accuracies achieved by ASR of bitstreams directly, *i.e.*, without decoding the CVSD bitstream in which case, the recognition accuracy is reasonably high, demonstrating that in applications like topic identification from large databases of speech, where computational time takes precedence over high accuracy, the bitstream based approach would be more suitable. The effect of preprocessing the speech signal for suppressing the quantization noise, is yet to be studied in the context of ASR for CVSD. We expect that the recognition performance would increase further on applying these methods to remove the uncorrelated noise from CVSD data.

References:

1. Lilly, B.T. and P.K. K. *Effect of Speech Coders on Speech Recognition Performance*. in *International Conference on Spoken Language Processing*. 1996.
2. Huerta, J. and R.M. Stern. *Speech Recognition from Speech Codec Parameters*. in *International Conference on Spoken Language Processing*. 1998.
3. Euler, S. and J. Zinke. *The Influence of Speech Coding Algorithms on Automatic Speech Recognition*. in *ICASSP-94*. 1994.
4. Sreenivasamurthy, N. and A. Ortega. *Towards Efficient and Scalable Speech Compression Schemes for Robust Speech Recognition Applications*. in *IEEE International Conference on Multimedia and Expo*. 2000.
5. Zheng, F. and J. Picone, *Robust Low Perplexity Voice Interfaces*. 2001, Institute for Signal and Information Processing.
6. Moreno, C.P., *Recognizing Voice over Ip: A Robust Front-End for Speech Recognition on the World Wide Web*. IEEE Transactions on Multimedia, 2001. **3**(2).
7. Kim, H.K. and R.V. Cox, *A Bitstream Based Front-End for Wireless Speech Recognition on Is-136 Communications System.* IEEE Transactions on Speech and Audio Processing, 2001. **9**(5).
8. Welch, V.C. *A Comparison of U.S. Government Standard Voice Coders*. in *Military Communications Conference, MILCOM '89*. 1989.
9. Steeneken, H.J.M., *Potentials of Speech and Language Technology Systems for Military Use: An Application and Technology-Oriented Survey*. 1996, NORTH ATLANTIC TREATY ORGANIZATION.
10. BlueTooth, *Bluetooth Standard Specification*.
11. Jayant, N.S. and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. 1984: Prentice-Hall Inc.