

Speech Recognition using Linear Discrimination Analysis Projected features obtained from GSM Codec Parameters

Juan Arturo Nolzco Flores

jnolzco@campus.mty.itesm.mx

Dpto. de Ciencias Computacionales, ITESM, campus Monterrey

Abstract:

The purpose of this work is to present some experiment results when an automatic speech recognition (ASR) system recognises GSM degraded speech. Acoustic models were trained either with 13 LP-cepstra coefficients obtained from the speech reconstructed for GSM coder or from 13 LDA (Linear Discrimination Analysis) projected features obtained from parameters of the GSM coder. The GSM database was obtained by artificially degrading the SPINE1 database. It was found that LP-cepstra coefficients obtained from the speech reconstructed for the coder was the most robust.

1. Introduction:

There are many tasks where an Automatic Speech Recogniser could be very useful, for example to replace the telephone operators or to help person when the eyes are busy. However, the large inter-speaker and intra-speaker variability made ASR also a very difficult task. At the beginning, the ASR attempted to recognise isolated words (i.e. phonemes or numbers)[1]. When the isolated word recognition accuracy was good enough, next step was to recognise continuous words (words separated by silences)[1] and the last step was to recognise continuous speech [1,2,3,4,5, 6, 7]. Other trend was also to make ASRs to work in real world, so the goal was to develop robust ASR systems. For example, robust to speech degraded for additive noise [8, 9] or robust to speech degraded for a communication protocol [10, 11].

In this work, we are concern with robust speech recognition for speech degraded by the coder of the GSM (Global System for Mobile communication) system. In section 2 a brief description of the cellular radio system is given, special emphasis is given to the GSM architecture and its speech coder. In section 3 are described the features used to make the ASR robust to GSM degraded speech. In section 4 some experimental results are given. Finally, section 6 presents some comments and conclusions.

2. Mobile Communication

Mobile radiotelephones were used for the maritime and military communication during the early decades of the 20th century [12]. In the 1950s CBs were installed in several cities. In the 1960s, an Improved Mobile Telephone System (IMTS) was installed. The advantage of this system is that each channel consists of a downlink frequency and uplink frequency. In the 1980s, an Advanced Mobile Phone System (AMPS) was installed. Since then, the Mobile communication is widely used, because they offer higher flexibility and mobility for the users and because companies have found that they need relative low infrastructure when compared with the infrastructure needed for Public Service Telephone Network (PSTN). In AMPS a geographic region is divided in cells (roughly circular), where each cell serves a limited number of users, and channel

frequencies are designed to maximise the number of users. At the center of each cell there is a base station (computer, transmitter/receiver and antenna), which receives the signals from the telephones inside the cell. To allow the communication between users from different cells, the base stations are also communicated between them.

In order to get connected, to transmit and receive the high quality speech signal between end users, the net needs communication protocols. The number of communications protocols and the way they are related is called the network architectures. There are a number of network architectures, for example GSM (Global System for Mobile Communication), IS-54 (TDMA-Time Division Multiple Access) and IS-95 (CDMA-Code Division Multiple Access) and IS-136 (TDMA-Time Division Multiple Access). Unfortunately, none of these systems is perfect and we need to deal with errors [12]. Moreover, since other goal of the Cellular Radio is to serve the maximum number of user, the speech has been coded (compressed) increasing this the speech distortion. In this work, we concentrate with speech degraded with the coder used in the GSM architecture.

2.1 Speech Coding

The speech can be coded using either waveform coding or parametric coding. The waveform coding tries to preserve the form of the wave and the parametric coders try to represent the speech with a set of parameters. The advantage of the parametric coder is that the speech is highly compressed. Some examples of waveform coders are PCM (Pulse Code Modulation), DPCM (Differential Pulse Code Modulation), and ADPCM (Adaptive Differential Pulse Code Modulation) [13]. The parameter coders are also divided depending the way they parameterise the speech, for example source-excitation coders or sub-band coding.

In the source-excitation model, a set of parameters are obtained from the source (bucal tract system) and other set of parameters are obtained form the excitation model (voice and unvoiced speech, in case of voice speech, the pitch is also needed). In most of the speech coders source parameters are the coefficients of the LP model (also know as LPC – Linear Prediction Coefficients).

Therefore, the main difference of the coders is the way they parameterise the excitation signal, some of them only use the pitch value if voiced signal is detected, others send an index of a codebook, others send only part of the excitation signal, some of them calculate an ideal excitation for the LP model. Some examples of parametric coders are LPC-10, RELP (Regular Pulse Excited Linear Prediction), CELP (Codebooke Exited Linear Prediction) and VELP (Vector Sum Excited Linear Prediction) [13].

2.1.1 Digital Cellular Radio: GSM

GSM is the European standard for mobile communication and it is used in more than 50 countries, inside and outside of Europe [12]. GSM uses a parametric speech coder called RPE-LTP (Regular Pulse Exited- Long Term Prediction) [14]. In the RPE-LTP the source parameters are the LPCs. In order to parametrise the residual, first the original speech is pass through an inverse LP filter. Once the residual is obtained, it is decimated 4 times (under-sampled every 4 samples). If regular under-sampling is used, then there are four possible under-sampled residual signals. RPE-LTP selects the under-sampled signal with more energy. With this parametrisation, in RPE-LTP the

speech is compressed eight times when compared with PCM (when speech is sampled at 8KHz and each sample has 13 bits).

In the receiver, as first step, the excitation residual signal is reconstructed (by interpolating the received down-sampled excitation signal). Next, the reconstructed excitation signal is pass through the LP model to generate the speech waveform.

3. Robust Speech Recognition:

When speech is degraded, then compensation can be done in different part of an ASR system. It can be done at preprocessing stage, it can be done at the acoustic model[15] or a combination of both [8, 9]. In this work, we develop compensation at preprocessing stage. We develop some experiments with three different features.

3.1 Robust Features for Acoustic Training

The speech data was coded using GSM [30]. The system was trained using three parallel feature streams for recognition (static parameters, first and second derivate). The feature stream used were:

- a) 13 LP-cepstra from the RPE-LTP reconstructed speech.
- b) 13 LP-cepstra obtained from the source parameters of the RPE-LTP coder in GSM.
- c) 13-dimensional Linear Discriminant Analysis (LDA) based projections were obtained from 26 features (13 obtained from the 13 LP-cepstra and 13 obtained from the residual cepstra).

4. Experiments and Results:

In this work, we used the SPINE1 database (Speech in Noisy Environments 1). This Coded Audio Corpus was created for the Department of Defense (DoD) Digital Voice Processing Consortium (DDVPC) by Arcon Corp., and produced by the Linguistic Data Consortium (LDC). There are a rough total of 19 hours and 28 minutes (~4.4Gb) of audio data.

Training data were approximately 30,000 sentences obtained from the SPINE1 training and test data. The testing was around 4000 sentences obtained from the SPINE1 test data. The acoustic models used 2,800 senons, each with 8 gaussian/state. Each HMM are two states, and the variance was not normalised. The RPE-LTP coded data were obtained by artificially degrading the SPINE1 database [14].

The baseline is obtained with uncoded speech. This baseline gives 23.702% error rate. When the speech is coded/uncoded with the RPE-LTP coder, the recognition rate is degraded to 26.301%. When the speech is coded with the RPE-LTP and the LP-cepstra are obtained directly from the LP parameters of the coder, the signal is degraded to 28.9%. This is understandable, because the information of the residual signal is not taken into account.

In the next experiment 13 LP-cepstra were calculated from the LP parameters of the RPE-LTP, and also 13 LP-cepstra were calculated from the residual signal. These 26 coefficients were projected to 13 coefficients using LDA and the results show a 29.057%

error rate. When further decorrelation, using Karhunen-Loeve transform [16], was applied the error rate decrease to 27.577. The results are summarized in the following table:

Condiciones de Prueba	% wer
uncoded speech :	23.702
gsm coded/decoded speech :	26.301
with lpc cespstral	28.900
with lpc cespstral+residual cep (13 dimensional, after LDA)	29.057
with lpc cespstral+residual cep (13 dimensional, after LDA)+further decorrelation of features	27.577

Tabla 1: Results

5. Comments and Conclusions

In this work are presented some experiment results when an automatic speech recognition (ASR) system recognises GSM degraded speech. Acoustic models were trained either with 13 LP-cepstra coefficients obtained from the speech reconstructed for GSM coder or from 13 LDA (Linear Discrimination Analysis) projected features obtained from parameters of the GSM coder. The GSM database was obtained by artificially degrading the SPINE1 database. It was found that LP-cepstra coefficients obtained from the speech reconstructed for the coder was the most robust.

As a future work, we need to understand more deeply the distortion generated for the coders such that we can generate new compensation techniques.

Acknowledges

This research was supported for "La Academia Mexicana de Ciencias" and "la Fundación México-Estados Unidos para la Ciencia".

Bibliografía

- [1] L. Rabiner & B-H Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.
- [2] F. Jelinek, Continuous Speech Recognition by Statistical methods, Proceedings of the IEEE, vol. 64, , pp 532-556, 1976.
- [3] S.J. Young, *et al.*, Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems, CUED/F-INFENG/TR.38, Julio, 1989.
- [4] L.R. Bahl, F. Jelinek y R.L. Mercer, A Maximum Likelihood Approach to Continuous Speech Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 5, pp. 179-190, Marzo, 1983.
- [5] K. Kita, T. Kawabata & H. Saito, HMM Continuous Speech Recognition using predictive LR parsing, Proceeding of the IEEE International Conferences on Acoustics, Speech and Signal Processing, Toronto, Canada, 1989.
- [6] K. Seymore, *et al.*, The 1997 CMU Sphinx-3 English Broadcast News Transcription System, Proceedings of the Broadcast News Transcription and Understanding workshop, February 1998.

- [7] F. Kubala, et al., The 1997 BBN Byblis System Applied to Broadcast News Transcription, Proceedings of the Broadcast News Transcription and Understanding workshop, February 1998.
- [8] J.A. Nolasco Flores, & Young, S.J., Continuous Speech Recognition in Noise using Spectral Subtraction and HMM Adaptation, Proc. ICASSP, Adelaide, Australia, Apr. 1994.
- [9] J.A. Nolasco Flores & Young, S.J., Adapting a HMM-based Recogniser for Noisy Speech Enhanced by Spectral Subtraction, Proc. EUROSPEECH, Berlin, Alemania, Sept. 1993.
- [10] J.M. Huerta, R.M. Stern, "Speech Recognition from GSM Codec Parameters", ICSLP 1998.
- [11] J.M. Huerta R.M. Stern "Distortion-class weighted acoustic modeling for Robust Speech Recognition under GSM RPE-LTP coding" Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions Tampere 1999.
- [12] A. S. Tanenbaum, Computer Networks, Prentice Hall, Third Edition, 1996.
- [13] D. O'Shaughnessy, Speech Communication: Human and Machine Addison Wesley series in Electrical Engineering: Digital Signal Processing, 1987.
- [14] Digital Speech Compression, Putting the GSM 06.10 RPE-LTP algorithm to work, <http://www.ddj.com/documents/s=1012/ddj9412b/9412b.htm>
- [15] M.J.F. Gales and S.J. Young, Robust Continuous Speech Recognition using Parallel Model Combination. IEEE Transactions on Speech and Audio Processing Volume 4, 1996.
- [16] Stanislav Gordeyev, Karhunen-Loève Decomposition: Literature Review, <http://www.nd.edu/~sgordeye/Project1/project1.html>