Very Low Bit Rate speech coding in Noisy Environments

M. Padellini[12], F. Capman[1], G. Baudoin[2].

(1) Thales Communications,
160, Bd de Valmy, BP82, 92704 Colombes, CEDEX (France)
(2) ESIEE ESYCOM, Telecommunications Systems Laboratory
BP99, 93162 Noisy-Le-Grand, CEDEX, (France)
marc.padellini@fr.thalesgroup.com

Abstract

Very low bit rate speech coders can offer high intelligibility at data rates below 800 bits/s. Many different schemes have been proposed during the last twenty years, but little attention has been paid to the effect of noise on speech quality. Yet, this tricky issue must be faced for real applications since most schemes use recognition models which are inherently sensitive to changes in the recording environment. In this article we study the effect of noise on Hidden Markov Models (HMM) and the selection of units in the context of very low bit rate speech coding. The speech coder considered reaches 500 bit/s on average, works online and is completely unsupervised. To improve its robustness to noise, different front-end speech features are compared as well as model adaptation and common spectral noise reduction techniques.

Introduction

Very low bit rate speech coders must use variable length segmentation to reduce their data bit rate without loss of intelligibility. Lately, several schemes have been proposed in [1], [2] and [3], which take advantage of a large speech corpus to code speech. The main shared idea is to build a segment classifier to handle segments instead of frames. But no study on the influence of noise on such coder has been made so far. Though bad recording conditions is a tricky issue, it must be faced to integrate this type of coder in real applications.

In this paper we propose to study the influence of noise on a VLBR coder which has been described in [4] and [5]. This coder operates at 500 bit/s and code speech online. In a recent evaluation described in [6] it has been ranked at the same intelligibility level than the NATO STANAG-4591 at 2400 bits/s.

After a short description of the VLBR coder, its noise sensitive parts is highlighted in Section 2. Different front-end speech features and noise reduction techniques are presented in Section 3, to improve the robustness of the HMM models. The influence of noise on the unit selection process is discussed in Section 4. Eventually, Section 5 presents the experimental tests and the results obtained under various configurations.

1. VLBR coding

The main idea in VLBR coding is to take advantage of a large speech corpus, looking up the elementary speech units that will best reconstruct the input speech. To achieve this goal, the VLBR coder combines two approaches: speech recognition using HMM in the encoder part, and synthesis by corpus in the decoder part. The speech database is built during a training phase.

1.1. Training phase

64 HMM models are used to jointly segment and classify the speech corpus. They are trained iteratively on the corpus using an initial transcription. This phase is completely automatic. Indeed, the segmentation is supplied by a dendrogram-like bottom-up frame merging process. The segments are classified using the cumulated distance to the code vectors computed by vector quantization of the speech corpus. The HMM topology is three states left-to-right models. The emitting probability of each state is modeled by one Gaussian. A Viterbi algorithm is used at the end of the training phase to jointly segment and classify the whole corpus. The units are gathered according to their class in the unit database.

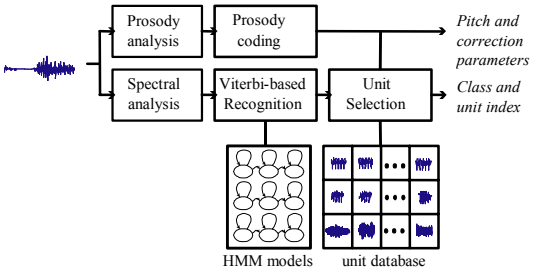

Figure 1: VLBR encoding principle.

1.2. Encoding phase

The encoding phase is presented on Figure 1. First, features are extracted every 10 ms. Pitch is computed using normalized cross-correlation.

1.2.1. Segment classification and unit selection

A Viterbi algorithm is used with previously trained HMM to derive from input speech the segmentation and the class of the target units. The unit selection is performed in two steps:

- Pre-selection: for each target unit, the 16 closest units of the same class in the synthesis database are pre-selected. The criteria used in the pre-selection is the mean pitch distance to the target unit.
- Final selection: normalized cross-correlations are computed between the target unit and the pre-selected units on their mean harmonic spectrum, energy and pitch profiles (see Figure 2). The unit which has the highest cumulated correlation is selected.

Two steps are used in the selection in order to code the selected unit with a fixed number of bits regardless to the

number of units in the synthesis database. Indeed, the class, the mean pitch and the index of the final selection are transmitted to retrieve the unit.

1.2.2. Unit correction

To best fit the target unit, the energy and pitch profile of the selected unit are corrected. The transformations are handled using an Harmonic plus Noise model (HNM Cf [7]).
An energy correction gain is computed to correct the mean level of the energy. It is defined by a ratio along the energy of the frames of the selected unit and the target unit. It is set in order to correct both the harmonic and stochastic part.
The pitch profile is corrected by a linear varying gain defined by the mean pitch of the target unit and a slope parameter (Cf [5]).

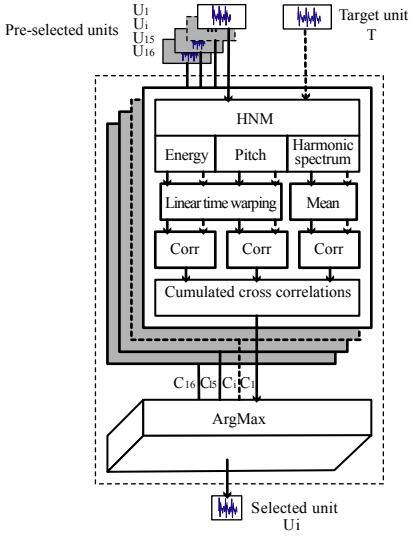

Figure 2: Final unit selection step

2. Recognition models in noise

Since the HMM models are trained on clean speech features, a mismatch between feature spaces is introduced when noise is added to the input speech. Segmentation and classification are no more consistent with the training. Changes in the class lead to the pre-selection of units that can't match the input speech.

2.1. Lack of transcription reference

The main problem in the evaluation of the robustness of the HMM is related to the type of unit considered in VLBR coding. There is no unambiguous transcription like phoneme transcriptions. For that reason, we will use as a reference the transcription obtained using the Viterbi algorithm on clean speech. This reference is specific to the type of features used to train the HMM.
Given the clean speech and the noisy speech transcriptions, recognition scores can be computed by a dynamic programming-based label alignment procedure, using the HRESULTS tool of HTK (Cf [8]). The percentage number of labels correctly recognized is given by:

$$\%Corr = \frac{H}{N} \cdot 100\% \quad (1)$$

Were H is the number of correct labels and N is the total number of labels in the transcription files.
The recognition scores can't be compared with classical phoneme recognition systems. Actually, under the effect of noise, one unit could be splitted into two units of different classes. Since there is no known optimal transcription, the resulting synthesized speech is not inevitably worse.
Furthermore, there is 64 classes of units: it is twice the number of classes of phoneme recognition systems. On one hand it lowers the recognition scores. But on the other hand the wrong recognized classes are closer acoustically which means the effect of errors on speech quality is reduced. This aspect suggests that a soft error decision should be considered in the evaluation. But in practice we will assume that the closest the transcription is to the clean speech transcription, the best is the noise robustness.

2.2. Reference cepstral features

A large number of features can be used in speech recognition, we tested in this article :

- LPCC parameters (Linear Predictive Coding Cepstrum). The cepstrum is derived from the LPC spectral envelope.
- MFCC parameters, (Mel Frequency Cepstrum Coefficient). The cepstrum is derived from Mel filterbank outputs. It has been recently standardized in the ETSI ES 202 050 standard (Cf [6]).

These features are not robust to noise since they try to model the spectrum without any noise assumption.

2.3. Tested methods for noise robustness

Noise reduction techniques can reduce the mismatch introduced by additive noise in the feature space. They can be applied at three different stages: before feature extraction, during feature extraction or on the trained models.

2.3.1. Speech enhancement

Noise can be subtracted in the spectral domain before feature extraction, using a noise model estimated on regions where speech is not active. A Voice Activity Detector (VAD) can be used to supply such regions. Then, a noise adaptive filter can be designed using MMSE (Minimum Mean Square Error Short Time Spectral Amplitude Estimator). It has been used successfully in [9] to enhance speech and can be used to derive noise robust features.

2.3.2. Joint Feature extraction and noise reduction

Lately, a feature extraction has been standardized in ETSI ES 202 212 (Cf [10]). Noise is estimated using a VAD, it is subtracted during feature extraction, using a two stage Wiener filter applied on Mel filter bank outputs. Robust MFCC are derived, which are called AURORA MFCC in the following.

2.3.3. Noise robust statistical modeling

Instead of trying to remove noise from speech, which can introduce strong distortions, noise can be added to the clean speech models. For that purpose, Parallel Model Combination (PMC) can be used. It has been studied by Gales in [11]. A HMM noise model is trained and combined to speech HMM models in the spectral domain using a Log-Add

approximation. Noise and speech spectral means are added. Resulting spectral means are then transformed back into the cepstrum domain.

3. Modifications for noise robust unit selection

In the last section we have considered the effect of noise on HMM models and more specifically on speech features. But it can be expected that noise can also impact on unit selection. For that reason, pitch estimation must be robust enough to have consistent pre-selection of the synthesis units.

3.1. Effect of noise on unit correction

Provided that pitch is robust enough, noise has little effect on final unit selection since the synthesis units are close to each other: they have the same class and close mean pitch. In fact, the main problem lie in the correction of the selected units. Two cases can be considered:

- Recognition failed. The pre-selected units are very different to the target unit. To match the target unit, strong gains are estimated to correct the energy and pitch profiles. It results in strong discontinuities. Nothing but improving the class recognition and limit corrections can be done.
- Recognition succeeded. Pre-selected synthesis units are good representatives of the target unit. Because of noise, the stochastic part of the target unit is strong and an overestimated energy correction gain is computed. Therefore, artificial noise is synthesized through the stochastic part of the HNM model.

3.2. Modified unit correction

To avoid artificial noise to be synthesized, two different approaches can be used which led to similar quality. Either remove noise from the target units using MMSE noise reduction or add noise to the synthesis units. These methods were tested by bypassing the recognition models and using the clean speech transcriptions to prevent the effect of bad class recognition (first case). In the latter method, noise was added using the mean HNM profile of the first second of the noisy speech, before the start of the utterance.

4. Experimental tests

4.1. Speech Material

Four speakers were took from the French speech corpus BREF (see [12]). These speakers are two males and two females. For each speaker, one hour of speech was divided into a training and a testing corpus. 10 test utterances were reserved for testing. All the signals used in the experiment were wide-band signals sampled at 16 000 Hz.

4.2. Noise Material

Three noises of 1 minute long were used in the experiments :

- “Subway”: Travel of a subway between two stations, including departure and arrival.
- “Car”: A Smart car at 80 km/h recorded on a ring road at a fairly stationary speed.
- “Babble”: canteen, 100 people. Taken from the NOISE-ROM-0 noises (Cf. [13])

Figure 1 shows the spectrum density of the noises.

4.3. Training

The training phase was performed on each speaker using the training corpus and using separately three features: LPCC, MFCC, and AURORA MFCC (as described in Section 2). We obtained respectively 3 sets of HMM. Each set is compound of 63 speech HMM and 1 silence HMM.
These HMM sets were used to derive three transcriptions for the test utterances. These transcriptions were used as references to compute recognition scores. They were also used to build a synthesis database for each type of feature.

4.4. Simulated noisy speech encoding

The SVP56 tool from the ITU-T Recommendation P.56 (Cf [14]) was used in root mean square mode to set the right levels of speech and noise in order to obtain noisy speech test utterances at a specific SNR. Noise regions were drawn randomly.
The three HMM sets trained previously were used to derive the transcriptions of the noisy speech test utterances at different SNR levels. The recognition scores were computed by comparing the clean speech and noisy speech transcription corresponding to the type of feature considered.
This process has been repeated using two noise adaptation techniques and only one type of feature:

- PMC adaptation with LPCC features. The noise models were trained using one second of noise before the start of the utterance.
- MMSE enhancement with MFCC features: This technique was used in front end of feature extraction.

4.5. Results

The mean recognition scores obtained are reported on Tables 1, 2 and 3 for each noise, feature and SNR considered. We can see there is little difference between LPCC and MFCC parameters, though LPCC are slightly better under Car and Babble noise.
The recognition scores of transcriptions obtained through noise reduction techniques show that AURORA MFCC features are more robust when compared to MMSE MFCC. At any rate the PMC LPCC gave the best recognition scores. Indeed, its insertion rate were very low since the silence model is better adapted.
In Figure 4 are plotted the spectrograms of a clean speech test utterance, its noisy version at 15dB SNR under car noise and its VLBR coded version. The unit correction was adapted adding noise to speech units (as described in section 3) and PMC LPCC were used. A good quality was obtained at this SNR even with car and babble noise, but when SNR reaches 5 dB the pitch extraction is not robust enough. The quality of the unit selection is poor and wrong pitch correction introduce strong discontinuities. Examples of this encoded utterance can be found for every noisy condition at:
www.esiee.fr/~baudoing/sympatex/demo

5. Conclusions

A study of different parameters and noise reduction techniques showed that VLBR coding of speech in noisy environments can be done using noise reduction techniques such as PMC. Good quality can be obtained under various

noises but the pitch estimation has to be improved to handle SNR under 15dB. A combination of the three noise reduction techniques could be studied since noise filtering usually performs better at very low SNR than PMC. Moreover class recognition could be improved by extending the unit selection to the classes which are close to the recognized class.

6. Acknowledgements

This work has been carried out under the grant n° 449/2002 from the ANRT (CIFRE) and RNRT project SYMPATEX.

7. References

[1] Cernocky J., Baudoin G., Chollet G., "Segmental vocoder - going beyond the phonetic approach," *Proc. ICASSP-98, pp. 605-608*, 1998.

[2] Lee K.S., Cox R., "A Segmental speech coder based on a concatenative TTS", *Speech Communication, Vol. 38*, *pp 89-100*, 2002.

[3] Ribeiro C.M., Trancoso I.M., "Phonetic vocoding with speaker adaptation", *Proc. Eurospeech*, *pp 1291-1294*, 1997.

[4] Baudoin G., Chami F.El, "Corpus based very low bit rate speech coding," *Proc. ICASSP-03,* pp. 792-795, 2003.

[5] Padellini M., Capman F., Baudoin G., "Very Low Bit Rate (VLBR) Speech coding around 500 bits/sec", *EUSIPCO*, Vienna, 2004.

[6] Padellini M., Capman F., Baudoin G., "Evaluation d'un codeur de parole à très bas débit", *to appear at GRETSI. 2005.*

[7] I. Stylianou, *Modèles Harmoniques plus Bruit combinés avec des Méthodes Statistiques, pour la Modification de la Parole et du Locuteur.* ENST PhD, 1996.

[8] Young S. and al. "The HTK Book", *Cambridge University Engineering Department*.

[9] Ephraim, Y., Malah D., "Speech enhancement using Minimum Mean Square Error Short Time Spectral Amplitude Estimator", *IEEE Trans. On Acoustics, Speech and Signal Processing, 32(6), pp. 1109-1121,* 1984.

[10] ES 202 212, "Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms;", *ETSI document*, August 2003.

[11] Gales, M.J, "Model-based techniques for noise robust speech recognition", *PhD thesis, University of Cambridge,* September 1995.

[12] Lamel L.F., Gauvain J.L., Eskenazi M., "BREF, a large vocabulary spoken corpus for French," *Proc. EUROSPEECH-91,* Genoa, Italy, 1991.

[13] "NOISE-ROM-0", *NATO:AC243/(Panel 3)/ RSG-10. ESPRIT Project n°2589-SAM.*

[14] "ITU-T Software Tool Library 2000 User's Manual", *ITU-T Users' Group on Software Tools*, Geneva, December 2000.

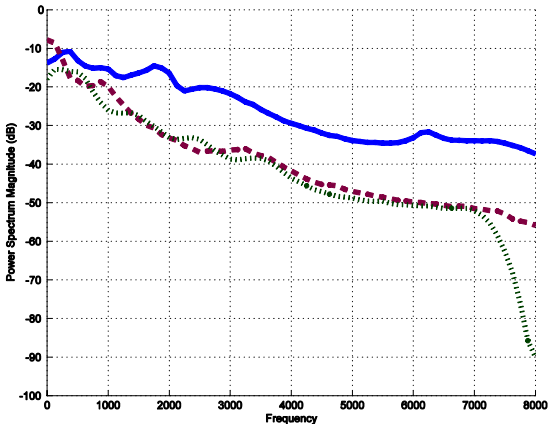

Figure 3: *Power spectrum of subway noise (solid line), Car Noise (dashed line), Babble noise (dotted line)*

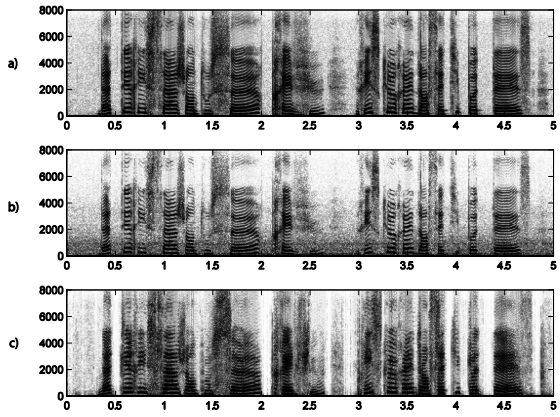

Figure 4: Spectrogram of original utterance (a), after adding car noise(15 dB SNR) (b), after VLBR coding using PMC LPCC features (c).

	Features	LPCC			MFCC			AURORA MFCC			MMSE MFCC			PMC LPCC		
	SNR(dB)	**5**	**15**	**20**	**5**	**15**	**20**	**5**	**15**	**20**	**5**	**15**	**20**	**5**	**15**	**20**
Subway	Corr (%)	15	36	45	19	41	47	34	53	57	29	51	52	35	51	62
	Sub (%)	82	59	50	76	53	46	59	39	36	65	42	41	41	36	29
	Ins (%)	26	28	23	29	23	22	27	21	21	26	22	21	5	10	10
Car	Corr (%)	31	57	61	30	55	58	45	61	64	41	53	55	51	72	75
	Sub (%)	64	39	34	64	39	36	48	32	29	52	40	38	31	18	16
	Ins (%)	34	22	21	26	22	21	22	17	17	26	22	22	6	5	5
Babble	Corr (%)	29	54	59	28	51	57	36	54	59	33	50	54	45	63	72
	Sub (%)	68	43	38	69	44	39	58	39	36	61	44	40	45	29	22
	Ins (%)	38	29	27	37	30	30	33	23	23	30	27	27	19	14	11

Table 1, 2, 3: Recognition rates under Subway, Car and Babble noise. (Corr: Correct, Sub: Substitution, Ins: Insertion)