PHONETIC VOCODER ASSESSMENT

Carlos M. Ribeiro INESC/ISEL-CEDET
Isabel M. Trancoso INESC/IST
Diamantino A. Caseiro INESC/IST

INESC, Rua Alves Redol, 9, 1000 Lisbon, Portugal.
E-mail: {cmr, Isabel.Trancoso}@inesc.pt Phone: +351 1 3100 314 Fax: +351 1 3145843

ABSTRACT

The efficiency of phonetic vocoders stems from the fact that the only transmitted information is the index of the recognised units and the corresponding prosodic parameters. Hence, speaker recognisability is one of the main issues in this class of coders. Our approach to minimise this drawback was to include some speaker adaptation capability. The purpose of this paper is two-folded: on one hand, to describe the recognisability and intelligibility tests that were performed with our phonetic vocoder with and without speaker adaptation; on the other hand, to present our recent developments of this coder, using the SpeechDat *corpus* for Portuguese, that includes telephone calls from 5000 speakers. This allowed us to generate improved HMM models, codebooks, and quantization tables, and to investigate the performance of the coder in non-clean environments and with a much wider speaker population.

1. INTRODUCTION

Phonetic vocoding [2] is one of the techniques able to reach very low bit rates. This efficiency stems from the fact that the only information transmitted by this type of coder is the index of the recognised unit and the corresponding prosodic parameters. Hence, speaker recognisability is one of the main issues in this class of coders. Our approach to minimise this drawback was to include some speaker adaptation capability in the coder [4] [5]. The present paper describes our recent work on this coder, involving on one hand its formal assessment in clean laboratory conditions and, on the other hand, its performance in non-clean environments with a much wider speaker population.

The formal assessment of the coder included two sets of tests: speaker recognisability tests designed to assess the potential advantages of the adaptation procedure, and intelligibility tests. No MOS tests were performed. In fact, these tests involve trained listeners, which are confronted with processed sentences and requested to judge the quality in a five-point scale (1-5). However, listeners normally avoid marking the quality in the scale extremes. This no linearity makes the MOS test useless for very low bit rate, as phonetic coders.

The version of the phonetic coder used in the two sets of tests was developed and tested using a telephone-bandwidth version of the EUROM.1 *corpus* for European Portuguese [3]. This *corpus* includes 60 speakers recorded in an anechoic room, but only 10 of them generate a reasonable amount of speech. The subset used for training both the HMM recogniser and the synthesis codebook included only 8 of these speakers, corresponding to 32 minutes of speech. A distinct subset was used for the above mentioned tests. The EUROM.1 *corpus* is obviously quite small to analyse intra and inter-speaker variability, a major issue in the development of this type of coders. More recently, the SpeechDat *corpus* for European Portuguese [9] became available, including telephone calls from 5000 speakers. This availability resulted in the generation of new HMM models, codebooks, and quantization tables, potentially allowing the study of the performance of the coder in non-clean environments and with a very wide speaker population. Formal recognisability and intelligibility tests have not yet been performed with this version.

The speaker adaptation procedure will be very briefly reviewed in section 2. The two following sections respectively describe the speaker recognisability and the intelligibility tests we performed. The recent developments with the SpeechDat *corpus* will be described in section 5. Finally, section 6 will address conclusions and further work.

2. PHONETIC CODER WITH SPEAKER ADAPTATION

Like in other basic LPC vocoders, the phonetic encoder performs LPC analysis, and estimates pitch, voicing and energy parameters, on a frame by frame basis. In our phonetic vocoder, however, the LPC coefficients are fed into a HMM phone recogniser to derive a phone index. The transmitted information is now the phone index and duration, together with pitch, voicing and energy information, thus resulting in a variable bit rate scheme with an average bit rate of 443 bit/s (computed for the EUROM.1 *corpus*).

In the decoder, the phone index together with the previous and next indexes, are used to retrieve a set of LSF coefficients from a context-dependent LSF codebook. Time scale modification was adopted to adjust the duration of the stored phone in the codebook to the transmitted duration. The restored LSF coefficients are then used, together with energy and pitch information, to produce the synthetic speech, using a conventional LPC vocoder scheme. Naturally, the synthetic speech is very dependent on the speakers used to generate the codebook.

The first attempt to adjust the synthetic signal to the input speaker was to include a gender identification model based on

the average pitch, which was already transmitted, in order to access a gender dependent codebook. This module achieves 95% correct identification, with a majority rule over the last 7 voiced segments.

To further adapt the codebook to the input speaker, a speaker adaptation scheme was developed, where the average values of each LSF coefficient, computed over the whole duration of the phone, must be transmitted. This extra speaker specific information is only transmitted for vowel and glide phones, where the speaker characteristics are perceptually more important. This information required, in average, only 116 bit/s, bringing the total bit rate to 559 bit/s. In the receiver, the LSF coefficients are restored to match the transmitted average values that carried speaker information, but relying on the dynamic characteristics stored in the codebook. The adaptation procedure is based on the minimisation of a spectral distance [5], following the expression:

$$LSF_{i_{mod}} = LSF_{i_{codebook}} - \overline{LSF}_{i_{codebook}} + \overline{LSF}_{i_{input}}, \quad (1)$$

where *codebook* stands for stored *LSF* values, *input* for transmitted values and *mod* for adapted values.

3. SPEAKER RECOGNISABILITY TEST

In order to test the potential advantages of the adaptation procedure, a speaker recognisability test was carried on. The test was based on the methodology used to select the new 2400 bit/s DoD speech coder [7][8], where pairs of utterances are presented to a listener to judge if they are spoken by the *same* or by *different* speakers. Two sets of experiments were conducted. The first experiment verified the degree to which each coder preserved speaker identity. Hence, the first sentence of the pair was not coded (U-P unprocessed-processed). The second experiment verified how well each coder preserved the information necessary to distinguish one speaker from another. Hence, the coder processed both sentences (P-P processed-processed). Additionally, the listeners rated the dissimilarity between the two voices using a six-point scale listed in Table 1. As [7] used a five-point scale, Table 1 shows the mapping we used between these two scales.

Dissimilarity		Meaning	5-point mapping
Same speaker	1	I´m positive it's the same speaker	0
	2	I think it's the same speaker	1
Ambivalence	3	I'm not sure, but it's probably the same speaker	2
	4	I'm not sure, but it's probably a different speaker	2
Different speakers	5	I think it's a different speaker	3
	6	I'm positive it's a different speaker	4

Table 1: Dissimilarity scale

The two versions of the phonetic vocoder, with and without speaker adaptation were assessed. As references, the FS1015 vocoder and the unprocessed speech were also assessed. We used 20 speakers, 10 female and 10 male. In order to reduce the test complexity, these speakers were grouped into 4 blocks of 5 speakers. 8 untrained listeners have performed the test.

The percentage of correct responses, shown in Table 2, was computed individually for *same* and *different* pairs in both experiments, using the area under the receiver operating characteristic, already used in [7]. It's generally easier to distinguish between *different* speakers than to identify the *same* speaker, even when judging unprocessed speech.

	U-P			P-P		
Coder	Avg.	Same	Diff.	Avg.	Same	Diff.
Unprocessed	86	81	91	----	----	----
Reference coder	69	52	85	80	73	88
phonetic coder with adaptation	66	54	78	74	70	78
phonetic coder without adaptation	59	35	83	66	67	65

Table 2: Percentage of correct responses for the *same* pairs, *different* pairs and average values.

The phonetic coder without speaker adaptation achieved a very low percentage of correct answers (35%) for the *same* speakers in the U-P experience. This score may be justified by the fact that the output speech results from a *typical* voice based on the speakers that contributed to the codebook generation, the main differences being on the prosodic aspects. For *different* speakers, the percentage of correct answers increased, as expected. The improvement introduced by the speaker adaptation procedure can be verified in all the test conditions, by comparing the results with and without speaker adaptation. The difference between the version with speaker adaptation and the reference coder is 3 (U-P conditions) and 6 (U-P conditions) points for the two experiments.

Similar values are listed in Table 3, but computed individually for each gender. On average, female speakers were harder to identify than male speakers. This was also verified for unprocessed speech.

	U-P			P-P		
Coder	Avg.	M	F	Avg.	M	F
Unprocessed	86	89	84	----	----	----
Reference coder	69	68	70	80	82	78
phonetic coder with adaptation	66	68	65	74	74	74
phonetic coder without adaptation	59	62	57	66	61	72

Table 3: Percentage of correct responses for each gender and the corresponding average value.

Average dissimilarity judgements are shown in Tables 4 and 5, computed individually for *same* and *different* pairs in both

experiments. The best coders had smaller dissimilarity judgements than the poor coders for the *same* pairs in unprocessed-processed conditions. For *different* pairs in processed-processed conditions, the best coders had larger dissimilarity judgements than for the poorest coders. These two statements are true both for correct and incorrect answers. The phonetic coder with speaker adaptation yielded better results than without speaker adaptation.

	U-P			
	Same		Different	
Coder	Corr.	Incorr.	Corr.	Incorr.
Unprocessed	0,36	3,13	3,72	1,16
Reference coder	1,06	3,08	3,28	1,29
phonetic coder with adaptation	1,15	2,82	2,97	1,52
phonetic coder without adaptation	1,28	2,96	3,13	1,40

Table 4: Dissimilarity judgement results for U-P conditions (five-point scale).

	P-P			
	Same		Different	
Coder	Corr.	Incorr.	Corr.	Incorr.
Reference coder	0,76	3,11	3,30	1,38
phonetic coder with adaptation	0,80	2,91	3,14	1,18
phonetic coder without adaptation	0,97	2,70	3,05	1,08

Table 5: Dissimilarity judgement results for P-P conditions (five-point scale).

After the test, listeners were informally asked about some generic quality aspects. No one showed any doubt about the speaker gender, even without speaker adaptation, validating the gender dependent codebook strategy and the automatic gender identification procedure.

4. INTELLIGIBILITY TEST

The DRT intelligibility test [1] is a very common assessment method for very low bit rate coders. It is a closed test where listeners choose between pairs of words with a /C-V-C/ structure and one different consonant (e.g., veal-feel). One of the disadvantages of this type of test is that it is not comparable between languages. An alternative is to perform an open test, where all the possibilities can take place in non-sense words, such as proposed in the SAM European project [6] for assessing synthesis systems, allowing the comparison of results from different languages. The test is limited to consonants in initial position (/CV/), central position (/VCV/) or final position (/VC/), combined with the three vowels /i/, /u/ or /ɑ/.

We have introduced several modifications to this procedure for assessing the intelligibility of the phonetic vocoder: the test segments were limited to central position (/CVCV/) in non-sense words, changing either the middle consonant or the vowel. The motivation for extending the test to vowels was the fact that these are the segments where speaker adaptation is performed.

Two sets of tests were performed: the first one assessed the intelligibility of 21 consonants or glides shown in Table 6 (presented in the /lɑ-*consonant or glide in test*+ɐ/ context); the second one assessed the intelligibility of 13 vowels shown in Table 7 (presented in the /t-*vowel in text*+tɐ/ context). The data was taken from two speakers, one male and one female, of the EUROM.1 CVC *corpus.*

p	t	k	b	d	g	f	s	ʃ	v	z	ʒ
ʁ	ɾ	l	ʎ	m	n	ɳ	j	w			

Table 6: Consonants and glides in test.

ɑ	ɛ	i	ɔ	u	ɐ	e	o	ɐ̃	ɛ̃	ĩ	ɔ̃	ũ

Table 7: Vowels in test.

The tests involved several versions of the phonetic coder, with and without speaker adaptation. In order to analyse the degradation due to recognition errors, versions with forced alignment were also included. As in the speaker recognisability test, the FS1015 vocoder and the unprocessed speech were also assessed as references.

The intelligibility rates, taking from judgements of 10 listeners, are presented at Table 8. These values can not be compared with DRT results, as this is an open test between 21 consonants or 13 vowels in non-sense words, and not between only two stimuli which are lexical words.

As far as vowel intelligibility is concerned, adaptation yielded an improvement of 20 percentile points (from 54% to 74%). As can be verified by the small difference of intelligibility scores obtained with speaker adaptation with (77%) and without (74%) automatic recognition, adaptation also helps solving recognition errors.

Coder	*consonants*	*vowels*
Unprocessed	94	99
Reference coder	82	95
Phonetic coder with speaker adaptation and forced alignment	50	77
Phonetic coder with speaker adaptation and automatic recognition	47	74
Phonetic coder without speaker adaptation and forced alignment	52	58
Phonetic coder without speaker adaptation and automatic recognition	36	54

Table 8: Intelligibility rate.

The intelligibility scores for consonants with forced alignment are almost the same with (50%) and without (52%) speaker adaptation, as no adaptation was made in this type of segments. With automatic recognition, however, the coder with adaptation performs better in consonants (47%) than without adaptation (36%), taking advantage of the better context.

5. EXPERIMENTS WITH THE SPEECHDAT CORPUS

The SpeechDat *corpus* was collected over two phases: a pilot phase with 1000 calls (SpeechDat I) and a main phase with the remaining 4000 calls (SpeechDat II). The subset used for training included only phonetically rich words and sentences from 80 % of the total set of speakers from both phases, maintaining the same proportion of age, gender and region distribution. The training set was first used for building continuous density hidden Markov models (CDHMMs). The feature extraction stage used conventional MFCC (14 cepstra + 14 delta-cepstra + energy + delta-energy), computed over 25 ms Hamming windows, updated every 10 ms. Gender dependent monophone models were initially built for 39 phones. The model topology was left-right, with 3 states, no skips, and 6 mixtures per state. Additional silence and filler models used forward and backward skips. After training monophone models, word internal tied state triphone models were trained, using tree based clustering. A total of 13k triphone models were built, with 8498 shared states. This recogniser achieved an accuracy of 90,7% on a development set, using a lexicon of 2356 phonetically rich words.

The generation of the LSF codebooks involved only the training set of SpeechDat II (close to 32 hours). Each generated codebook (male and female) included approximately 15000 different triphones (intra and inter-word). The previous EUROM.1-based version had only 4500 triphones. This reduces the substitution of non-existent contexts, which can introduce discontinuities in the synthetic signal.

Huffman codes were generated for encoding the phone sequence and duration. The total average bit rate, computed over the set of phonetically rich sentences from the training part of SpeechDat I, is 665 bit/s. This corresponds to 11 hours and 17 minutes of speech after removing silences, with a rate of 13 phone/s. The frame length, as in the HMM recogniser, is 10 ms instead of 11,25 ms of the EUROM.1 version, which slightly increases the bit rate.

The first set of experiments was done using forced alignment in order not to take into account recognition errors and better assess the influence of the environment. Informal listening tests confirmed the advantages of speaker adaptation and showed the need for further tuning of the coder in terms of adjusting the duration of plosives, which in this version were not separately modelled as closures and bursts. They also showed the noise reduction effect of substituting segments recognised as “filler” noises (typically speaker noises, environmental or channel noises of stationary or intermittent nature) by silence. The second set of experiments involving automatic recognition showed the advantages of the adaptation procedure in terms of recovering from recognition errors.

6. CONCLUSIONS

The first part of this paper presented the set of tests conducted to assess our phonetic coder with and without speaker adaptation. The speaker recognisability test showed an increase of 7 (U-P conditions) and 8 (P-P conditions) perceptual points, between the versions with and without speaker adaptation. The difference between the version with speaker adaptation and the reference coder is 3 (U-P conditions) and 6 (P-P conditions).

An intelligibility test was also conducted, which is an open choice test for consonants and vowels. By comparing the versions with and without speaker adaptation, an improvement of 11 points was obtained for consonants and 20 points for vowels.

The second part of this paper described our recent version of the phonetic coder based on the SpeechDat *corpus*. The availability of a much larger and realistic *corpus* allowed us to investigate some of the problems related with the presence of noise. It also opens the way to interesting experiments with the goal of reducing the speaker-specific transmitted information by trying to separately model intra and inter-speaker variability. We are also planning to investigate the issue of language dependency in this type of coder.

7. REFERENCES

1. J. R. Deller Jr, J. G. Proakis, J. H. L. Hansen, “Discrete-Time Processing of Speech Signals”, *Macmillan*, 1993.

2. J. Picone, G. Doddington, “A Phonetic Vocoder”, *Proc. ICASSP*, pp.580-583, 1989.

3. C. M. Ribeiro, I. M. Trancoso, M. C. Viana, “EUROM.1 Portuguese Database”, *ESPRIT 6819 SAM-A – Speech Technology Assessment in Multilingual Applications, Report D6*, 1993.

4. C. M. Ribeiro, I. M. Trancoso, “Improving Speaker Recognisability in Phonetic Vocoders”, *Proc. ICSLP*, 1998.

5. C. M. Ribeiro, I. M. Trancoso, “Speaker Adaptation in a Phonetic Vocoding Environment”, *Proc. of the 1999 IEEE Workshop on Speech Coding, pp.64-66*, 1999.

6. “SAM Final Report", *Section II - Speech Output Assessment*, 1992.

7. A. Schmidt-Nielsen, “A test of speaker recognition using human listeners”, *Proc. of the IEEE Workshop on Speech Coding for Telecommunications*, 1995.

8. A. Schmidt-Nielsen, D. P. Brock, “Speaker Recognizability Testing For Voice Coders”, *Proc. ICASSP, pp.1149-1152*, 1996.

9. I. Trancoso, L. Oliveira, “Portuguese Database for the fixed telephone network”, *Final Report*, 1998.