A codebook design method for fricative enhancement in Artificial Bandwidth Extension

Michele Sanna
Dipartimento di Ingengeria Elettrica ed Eletronica
Università degli Studi di Cagliari
Piazza d'Armi
09123, Cagliari, Italy
michele.sanna@diee.unica.it

Maurizio Murroni
Dipartimento di Ingengeria Elettrica ed Eletronica
Università degli Studi di Cagliari
Piazza d'Armi
09123, Cagliari, Italy
murroni@diee.unica.it

ABSTRACT

In mobile communications the transmitted speech signals are narrowband, thus sampled at 8 kHz. They are low-pass filtered under 4 kHz and a lot of intelligibility is lost. The goal of Artificial Bandwidth Extension (ABWE) is to recover the lost quality by reconstruction of the voice spectrum between 4 and 8 kHz, bringing thus the superior listening quality and intelligibility of wideband speech. The validity of an algorithm based on a Hidden Markov Model (HMM) has been demonstrated in the majority of speech variety, but resulted quite ineffective in the reconstruction of the fricative consonants. We investigated the causes of inefficient extension of the fricatives and the deriving problems. We developed a codebook design technique which provides a particular emphasis on these sounds in order to improve the fidelity of the reproduction and the dynamic of the processing. Our design improves noticeably the intelligibility of the fricatives. Log-spectral distance measures demonstrate the faithful extension as well as the subjective listening quality and intelligibility.

Keywords

ABWE, artificial bandwidth extension, speech enhancement, wideband speech, fricative

1. INTRODUCTION

Mobile telephony has still a low intelligibility due to narrowband speech coding. Speech coding was established to 8 kHz sampling PCM in 1972 (ITU-T G.711) as a compromise between bitrate and listening quality. Until today almost 100% of the infrastructure and devices still works with the narrowband standard. In order to sample at 8 kHz, speech signals are filtered in the bandwidth between 0.3 and 3.4 kHz, which ensures an intelligibility measured as the 90% of Syllable Articulation (i.e. the percentage of consonant-vowel-consonant words where all 3 component sounds are correctly recognized). Some listening quality is thus lost which is perceived as an imperfect comprehension.

The so-called wideband standard has instead a spectrum extended from 0.05 to 7 kHz and it is sampled at 16 kHz. It has a Syllable Articulation if 98% which means excellent intelligibility and clearness. The demand for wideband quality is constantly growing for hands-free devices, where the immediate understandability is critical to avoid the user's distraction, and for business mobile users with superior quality claims. In order to meet the demand, a wideband codec has been standardized for 3rd generation networks (AMR-WB, [4]), twenty years after the first wideband codec for ISDN (G.722), symptom of the necessity of a better conception of mobile communications.

Besides, coding is always more efficient and the transmission less onerous, due to the increased availability of pure bit-rate for mobile terminals. Unfortunately the substitution of several infrastructural devices, upon which 3rd generation networks still rely (like national and international backbones, regional backhaul links and the interfaces to the Public Switched Telephone Network (PSTN)) implies a conspicuous financial investment. Artificial wideband extension (ABWE) consists in estimating the part of the speech in the frequency band between 4 and 8 kHz, synthesizing speech with 8 kHz bandwidth. The estimation is made from normal narrowband speech, thus without the will of the sender to transmit any information about the missing band.

The algorithm makes 3G mobile devices receive wideband speech from a narrowband terminal or from another 3G user, connected by an old narrowband network. It can be placed at the interface between the two networks, at the base stations or even directly into the mobile terminal, making the conversation transparently wideband for the receiving user, even if it is carried over a narrowband network.

The source-filter technique for wideband extension was described in [6]. The power spectral envelope of a vocal tract filter is estimated in the missing band to shape a vocal tract filter, which is excited by a noisy wideband signal in order to synthesize wideband speech. An HMM-based system proved to have satisfactory performances with four different European languages (English, Spanish, French and German) and with cross-language training ([1, 2]). A wider interest is focusing on the improvement of the algorithm itself in prospect of a practical implementation for mobile and hands-free terminals.

Our study focuses on the performances of the algorithm

MobiMedia'09 September 7-9, 2009, London, UK

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7423
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7423

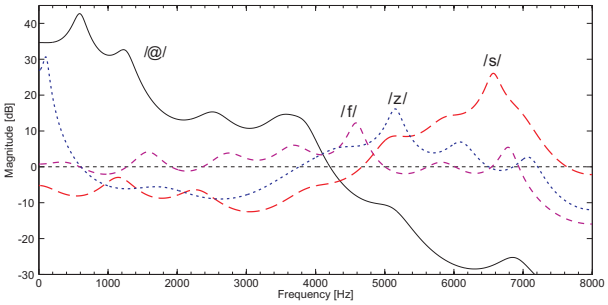

Figure 1: Confrontation between the spectra of three fricative consonants and of a typical vowel

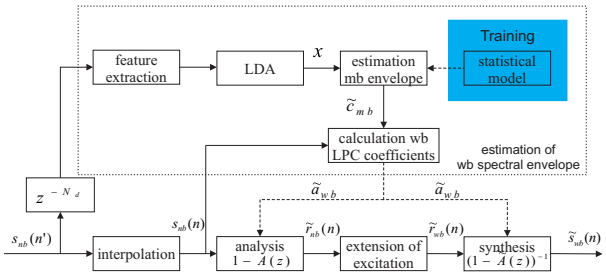

Figure 2: Block diagram of the artificial bandwidth extension (ABWE) algorithm

when dealing with particular phoneme classes. *Fricative* sounds such as /s/, /z/ and /f/, for example, are less effectively extended than other speech sounds, as it has been noticed also in [1]. The reason lies in their spectral shape. The energy is equally distributed between lower and upper band, or some times it is even higher in the missing band (Fig. 1). They are thus more affected than the other sounds by the lowpass filtering. They could receive appreciable benefits from the artificial extension, much more than other sounds that result already quite clear in the narrowband (e.g. the vowels). Unfortunately the low energy in the baseband makes the fricative sounds less detectable, the extension results more difficult and hardly definable with correctness. Later on this paper, Fig. 7 shows a comparison between a mild extension and an acceptable one in order to demonstrate the problem of underestimation. Paradoxically but not illogically fricative phonemes are those which could benefit much more from the extension but they are also the worst extended ones.

Since the core of the algorithm is the codebook, we focus the study on its expressiveness range and the space dedicated to each sound. The fricative sounds are only a few percentage of natural speech, so we propose a codebook design criterion and a training strategy that augment their statistical representation and thus the prospected presence in the speech. The main points of the design regard the enlargement of the envelopes codebook by adding some special states with high upper band energy and their aimed training to augment the weight in the estimation. We test the algorithm with the modified codebook using a database without transcriptions, thus simulating a telephony application which cannot rely on supplementary information. Cepstral distance measures demonstrate the faithful extension and the subjective listening experience confirmed the newfound intelligibility and quality.

After resuming the ABWE algorithm (Section 2) and the classic training strategies (Section 3) we explain the fricative oriented codebook design in Section 3.2. Section 4 shows the results of the log-spectral distance measures whereas Section 5 concludes the paper.

2. THE ABWE ALGORITHM

Artificial bandwidth extension based on Hidden Markov Models (HMM) was defined in [6]. The block diagram of the algorithm is shown in Fig. 2.

In order to restore the energy components in the missing band, the algorithm estimates, frame-by-frame, the power spectral envelope in the upper frequency range (called from now on the missing or upper band, in contrast to the lower or baseband, which is from 0 to 4 kHz). This step of the processing is performed by the upper blocks, obtaining the ceptral vector $\tilde{\mathbf{c}}_{mb}$. The estimated shape determines the frequency response of the complementary analysis-synthesis filters (lower blocks) in terms of the wideband filter coefficients $\tilde{\mathbf{a}}_{\mathrm{wb}}$. The analysis filter extracts a narrowband residual signal $\tilde{r}_{\mathrm{nb}}(n)$ from the input speech. After its wideband spectral extension, $\tilde{r}_{\mathrm{wb}}(n)$ excites the synthesis filter to produce wideband speech with the desired spectral shape.

Note: In the following, the terms "residual" and "excitation" are referred to the same signal. When the latter is used we are addressing it specifically as the input of the vocal tract (synthesis) filter, whereas the first has a more general validity.

2.1 Estimation of the envelope

In this stage the algorithm works on 10 ms frames of the narrowband signal $s_{\mathrm{nb}}(n')$, which is an admissible interval of stationarity. For every frame the upper band spectral envelope has to be estimated. The estimation goes through 3 stages:

2.1.1 *Feature extraction*

A 15-dimensional vector of features is extracted from the narrowband speech $s_{\mathrm{nb}}(n')$ according to [6]. The feature set includes 10 autocorrelation coefficients, the zero-crossing rate, the gradient index, the local kurtosis, the spectral centroid and the normed relative energy. The block of linear discriminant analysis (LDA) aims to reduce the feature vector to another 5-dimensional vector $\mathbf{x}$ which keeps valid the discrimination properties but reduces the dimension of the system and thus the computational effort.

2.1.2 *HMM, Hidden Markov Model*

The hidden Markov model is often used in the field of speech recognition [8]. It consists in N_s states, each one representing a possible upper band power spectral envelope. Each one of these envelopes is preliminarily trained and stored as a codebook class. By associating states and envelopes, the algorithm can calculate and assign state and transition probabilities ($P(S_i)$ and $P(S_i|S_j)$) for each state $S_1 \dots S_{N_s}$ of the Markov model. The observation probability density functions (pdfs) $p(\mathbf{x}|S_i)$ are described by a Gaussian Mixture Model (GMM) with pre-trained parameters and they are also assigned to each state. All the information that determines the probabilities and the pdfs is

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7423
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7423

extracted from natural speech, processed in the preliminary training stage and stored in the training set.

From the acquired state of the Markov chain at frame $m-1$ and the features extracted from frame m we can determine the condition of the Markov chain at the instant of the m-th frame, calculating the a posteriori probabilities by the Bayes formula [6, 1]:

$$P\big(S_i(m)|\mathbf{X}(m)\big) = C \cdot p\big(\mathbf{x}(m)|S_i(m)\big) \qquad (1)$$
$$\cdot \sum_{j=1}^{N_\mathrm{s}} P\big(S_i(m)|S_j(m-1)\big) \cdot P\big(S_j(m-1)|\mathbf{X}(m-1)\big),$$

where $\mathbf{X}(m) = \mathbf{x}(m), \ldots, \mathbf{x}(1)$ and C is a normalization factor. At the first frame we initialize using the state probability $P(S_i(m))$ in place of the summation.

2.1.3 *Calculation of the filter coefficients*

The algorithm choses the missing band envelope from the ones available in the codebook. A largely validated choice criterion is the Minimum Mean Square Error (MMSE) rule ([5]) which weights each codebook envelope (represented in the cepstral domain $\hat{\mathbf{c}}_{\mathrm{mb},i}$) by the a posteriori probability associated to its state and sums them up in the following way:

$$\tilde{\mathbf{c}}_{\mathrm{mb,MMSE}}(m) = \sum_{i=1}^{N_\mathrm{s}} \hat{\mathbf{c}}_{\mathrm{mb},i} P\big(S_i(m)|\mathbf{X}(m)\big). \qquad (2)$$

Other rules have been proposed in [6] and between them the Maximum A Posteriori (MAP) estimation can be still useful in place of the MMSE, in an implementation of the system that needs a particularly net decision. The MAP rule simply chooses the envelope S_{i_MAP} whose a posteriori probability $P(S_{i_\mathrm{MAP}}(m)|\mathbf{X}(m))$ is the biggest.

The estimated envelope is then transformed from the cepstral domain ($\tilde{\mathbf{c}}_\mathrm{mb}$) to the frequency domain by the inverse definition of cepstrum ([6]). Afterwards the upper band envelope is assembled with the known baseband power spectrum (periodogram of $s_\mathrm{nb}(n)$) to form a wideband envelope for the m-th speech frame. The coefficients $\tilde{\mathbf{a}}_\mathbf{wb}$ of the Linear Prediction (LP) filter are calculated from the assembled spectrum by the auto-correlation method ([6]). A fundamental property of the filter with such coefficients is that its frequency response has the same shape of the previously assembled wideband envelope.

2.2 Speech synthesis

In the source-filter method we need a cascade of two complementary filters, with inverse transfer functions: The analysis or Auto-Regressive (AR) filter $1-A(z)$ and the synthesis or Linear Prediction (LP) filter $(1-A(z))^{-1}$. The LP filter acts as vocal tract filter: It reproduces wideband speech from a noise-like wideband excitation signal.

To obtain such signal first the narrowband residual $\tilde{r}_\mathrm{nb}(n)$ is calculated by filtering the input speech $s_\mathrm{nb}(n)$ through the analysis filter (for coherence with linear prediction theory it performs a linear predictive coding, LPC by auto-correlation method [6]).

The missing part of the excitation signal is obtained by mirroring in the frequency domain the spectrum of the narrowband residual signal $\tilde{r}_\mathrm{nb}(n)$, which is by definition spectrally white, obtaining an 8-kHz-wide noisy signal $\tilde{r}_\mathrm{wb}(n)$ used as excitation [6]. This criterion does not affect the intelligibility of the speech because the human hearing is insensible to the fine structure at high frequencies [7] as well as the "hole" between 3.4 and 4.6 kHz which is not heard as a quality defect. The resulting speech $\tilde{r}_\mathrm{wb}(n)$ is wideband because of the extended residual and has the desired spectrum because the vocal tract filter has been designed to reproduce speech with a defined spectral shape by the estimated coefficients $\tilde{\mathbf{a}}_\mathrm{wb}$.

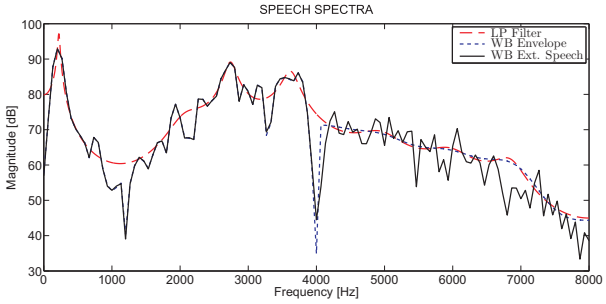

Figure 3: Spectrum, missing band envelope and LP filter frequency response of an artificially bandwidth extended signal

Table 1: Parameters of test and training systems

Par.	Description:	Value:
$f_{s'}$	sampling rate (NB)	8 kHz
f_s	(WB)	16 kHz
	FIR lowpass filter	
	order	252
	B_{3dB}	3.750 kHz
	Stopband attenuation	100 dB
N	Frame length in samples (10 ms)	160
N_-	Nř of preceding samples	40
N_+	Look-ahead	40
N_w	Total frame length	240
N_P	LP filter order (WB)	16
$N_\mathrm{a,bb}$	LP filter order (BB)	10
N_S	number of HMM states	
	original codebook	16
	new codebooks	24
G	GMM order	8
N_feature	number of features	15
y_flag	spectral envelope estimation rule	MMSE
		MAP
β	reduced feature dimension	5

Examples of spectra involved in the estimation are shown in Fig. 3. Notice that the filter transfer function follows the assembled envelope (compatibly with the filter order) and that the speech spectrum follows the filter response.

The parameters used in all our processings are resumed in Table 1.

3. SPEECH REPRESENTATION

3.1 Training

The estimation is driven by a statistical model that reproduces natural speech waveforms. In order to do that as much naturally as possible the algorithm is trained for the calculation of the model's parameters the training phase determines experimentally the right algorithm behavior by

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7423
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7423

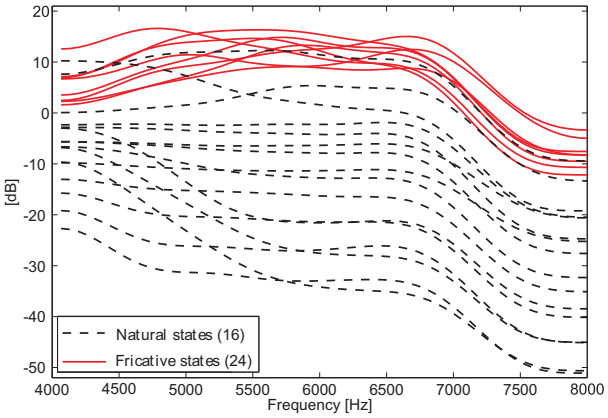

Figure 4: Extended codebook with 24 spectral envelope entries.

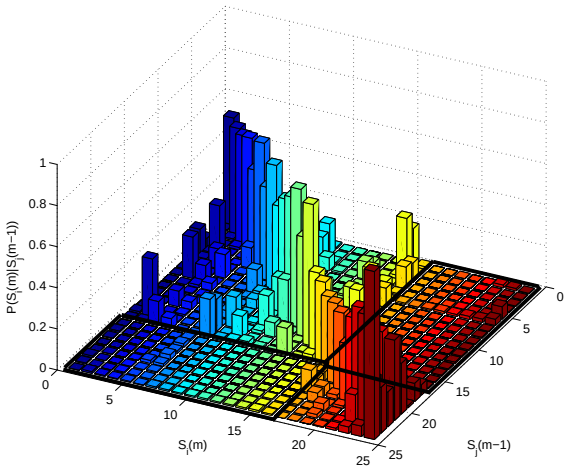

Figure 5: HMM Transition matrix after the labeled training.

observing and processing true wideband speech signals. The more training material is available, the best representation we have of natural speech in formal terms of the parameters listed here:

- A codebook of N_s cepstral vectors,
- A vector of state probabilities $P(S_i)$,
- A matrix of transition probabilities $P(S_i|S_j)$,
- A $(G + 2(G \cdot \beta)) \times N_s$ matrix of GMM parameters
- An **H** matrix of LDA transformation.

3.1.1 *Training of the VQ codebook*

The whole training material is processed through a Selective Linear Prediction (SLP) analysis that extracts from each frame a true cepstral representation of the upper band power spectral envelope. Then a Linde, Buzo, Gray (LBG) algorithm for vector quantizer design extracts 16 representative states that are saved into the codebook. These first states are called from now on the *natural* states because they are trained by vector quantization of the whole natural speech material ([1]). We also train in this phase the **H** matrix of the LDA.

3.1.2 *Training of the state model*

In this part the state's statistical parameters are determined. The upper band envelopes of the speech material are assigned to the codebook's entries by quantization.

Then the state and transition probabilities ($P(S_i)$ and $P(S_i|S_j)$) are calculated from the occurrences of state i and of couples of state j followed by an i one. Besides, the correspondences in a speech frame between the narrowband features and the quantized envelope concur to determine the observation probability density function of each state ($p(\mathbf{x}|S_i)$) in terms the Gaussian Mixture Model's (GMM) parameters [3].

3.2 Fricative oriented training

Previous works showed that the artificial extension of fricative sounds is quite difficult [1]. A weak point was found in the reproduction of fricative sounds. They are the most affected by the lowpass filtering because they can loose until 95% on the energy (for the sharpest fricatives we found), leaving in the base band a sort of white background noise that can be confused with speech pause. This loss of energy affects mainly the intelligibility because the nature of the sound is distorted for the human hear, but also for the algorithm which easily confuses the sound. In the technical system the misrecognition leads to a poor extension that actually does not restore the nature of the fricative, leaving the listening quality as poor as before (see Fig. 7 for comparison). Beyond the listening experience that revealed a leak in the audible quality, a local worsening of log-spectral distortion (LSD, [5]) was also verified during the measures, due to the too low estimated envelope.

We accept the fact that the chosen feature extraction method cannot overcome the necessary discrimination, due to the compromised nature of the sound. We adopt an ad-hoc training strategy to increment the attitude of the codebook, especially in its statistical representation, towards sharps sounds.

3.2.1 *Extension of the codebook*

In a naturally trained codebook, fricative sounds are insufficiently represented to permit a satisfactory extension. An arbitrary increase of the envelope, by constant factor or by fixed shape, is able give back to the fricative sound the right sharpness to be better intelligible.

In our design we increase the number of envelopes with high energy by training 8 special states from speech material composed only by fricative sounds. The material is vector quantized (after SLP) running the LBG algorithm with a large number of points and the states with the highest energy extrapolated and attached to the natural ones. The new states are called *fricative* states in contrast to the *natural* ones.

The goal of the enlargement of the codebook is to increase the representation of the fricative that in natural speech is statistically less relevant, and thus the presence in the final speech. The codebook with all the 24 states is shown in Fig. 4 were we can distinguish the fricative envelopes and the natural ones.

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7423
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7423

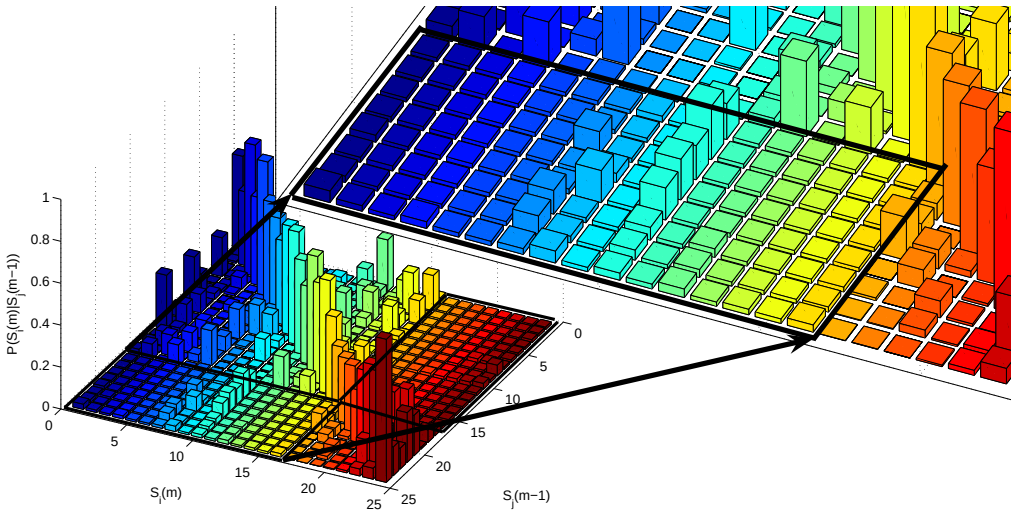

Figure 6: Detail of the HMM transition matrix after the retouch with $a = 0.01$ and $b = 0$.

3.2.2 *Transcription driven state assignment*

We need an ad-hoc procedure to deploy properly the new fricative classes of the codebook. Natural training applied to the extended codebook already demonstrates a good behavior. The new envelopes are quite involved in the state assignment, but of course the representation dedicated before by only two states to the fricatives (which is a minority set) is now divided between 8 states, each one only poorly deployed. Some slight improvement is verified, but the good potential is not yet being exploited.

We have the possibility of assigning selectively the fricative states by segmenting the training material. A labeled database is available for the training so that a distinction between fricative and non-fricative speech frames can be made. The training of the statistics is separated for the two sets of classes. During the assignment, the frames that contain a fricative sound are assigned to the fricative envelopes and concur to build their statistics about state and transition probabilities and the observation pdfs, exactly as explained before for natural training.

The segmented assignment is quite critical because we can compromise the result of the estimation process. The effects of segmented training are immediately noticeable in the state statistics. The involvement of the new states has been increased.

More consequent fricative sound frames train a lot of transitions between the 8 fricative states rather than from and to them, because during the sequence of fricative sounds all its frames are assigned to the fricative classes. The transitions from and to the fricative group are excluded, because it is excluded any occurrence of the normal states. Only two regular transitions are trained at the beginning and at the end of the fricative utterance, which is a quite low number compared to an amount of 10-15 processed frames.

The resulting structure for the HMM transition matrix is shown in Fig. 5. The influence is quite critical because it affects the estimation by Eq. 1. In the two marked rectangles we notice that very few value has is dedicated to the transitions from and to the fricative states, whereas high values are verified in the square of the new classes (17-24 $\times$ 17-24). With such configuration the inclination to avoid the switch from and to high envelope sounds is replicated during the technical system and some problems of latent dynamic might appear.

3.2.3 *Transition probability thresholding*

The training of the transitions suffers from lack of material. The 8 fricative states bring 192 new transitions. 64 of them are cyclically from a fricative state to another one so that they are already increased with the available of material. The other 128 transitions (marked rectangles in Fig. 5) can be hardly exploited, as explained before, so that a lot of values remain zero.

We use a criterion of compensation of lack of training material adding an arbitrary correction, according to the following law [9]:

$$P\big(S_i(m)|S_j(m-1)\big) = \\ = \begin{cases} C_j \cdot \Big(a + b \cdot P\big(S_i|S_j\big)\Big) & \text{marked rectangles} \\ C_j \cdot P\big(S_i|S_j\big) & \text{else,} \end{cases} \tag{3}$$

where C_j normalizes along the i dimension and the two constants a and b are chosen reasonably. A proportional correction can be applied, choosing $a = 0$, or a static one with $b = 0$. An intermediate solution is also possible. We implemented a static threshold of $a = 0.01$ with $b = 0$ to bring every value to an acceptable minimum and thus involve every possible state. In Fig. 6 is shown the increased probability floor that makes possible and reasonably probable every transition inside the marked rectangles.

With the expedient of thresholding we increase the time-dynamic of the estimation, i.e. the ability of switching quickly and properly from and to the fricative states, using the whole set of natural classes. The codebook is thus ready to provide envelopes with high energy, according to the envelope training, and the HMM has the right aggressiveness to place them correctly at the right moment. A slight affection of the other sounds is also expected, but the loss of quality is negligible compared to the gain that we suppose to verify for the fricatives.

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7423
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7423

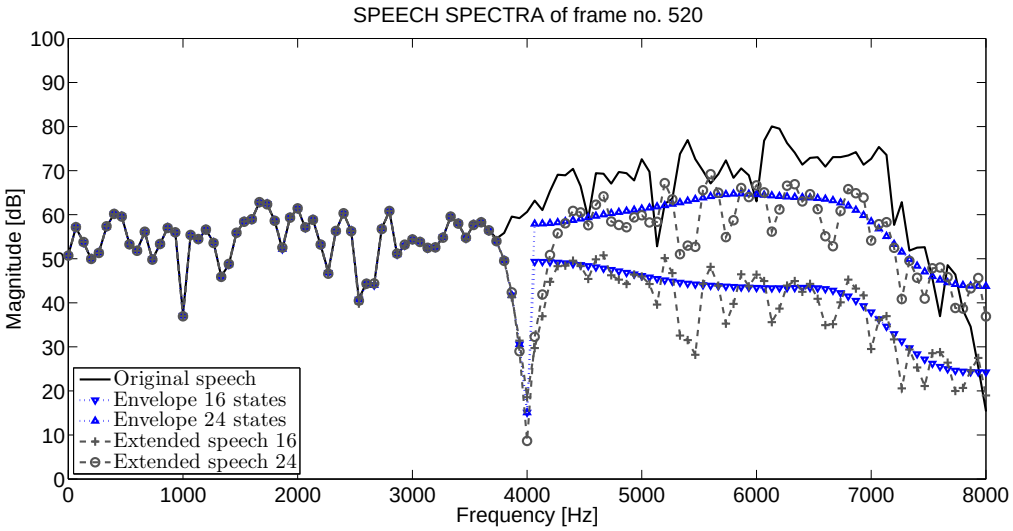

Figure 7: Comparison between two processing of the same speech frame using the old codebook and the new assembled one with segmented training and modified HMM.

4. RESULTS

4.1 Log-spectral distortion (LSD) metric

Extensive listening tests should be conducted in order to evaluate the gain in intelligibility and subjective quality. Unfortunately establishing a MOS session (Mean Opinion Score, ITU-T P.800) is impractical and the PESQ methodology (ITU-T P.862) to calculate the MOS score demonstrated quite untenable with extended speech, in contrast to the listening experience. We rely on the cepstral distance measure (LSD, log-spectral distortion) to evaluate the performances, as already used in [5, 1]:

$$d_{\mathrm{LSD}} = \frac{10}{\ln 10} \cdot \sqrt{(c_0 - \tilde{c}_0)^2 + 2\sum_{d=1}^{\infty}(c_d - \tilde{c}_d)^2}, \qquad (4)$$

where the terms of confrontation are the upper band cepstral representation $\mathbf{c}_{\mathrm{mb}}c$ of the original and estimated signals' envelopes. The index of performance is thus that the lower value of distortion is measured, the closer the estimated envelop is to the original one.

4.2 LSD measures

The measures are carried out over a database of multi-language speech signals, different from the database used during the training to avoid correlation. There are no transcriptions available thus the tests simulated an on-line telephony application.

A local improvement of the LSD is verified in correspondence to the fricative consonants, as shown in Fig. 8, were the measures of the following 4 experiments are plotted:

1. 16 states codebook,
2. 24 states codebook with segmented training,
3. 24 states codebook with natural training,
4. 24 states codebook with segmented training and HMM retouch.

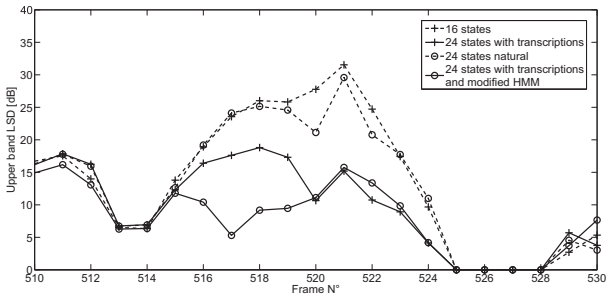

Figure 8: Upper band LSD measures using 4 codebooks. Detail relative to the /s/ phoneme in the utterance *"first"* (Frames 516-523).

Even if the new codebook performs slightly better already with natural training (3rd experiment), the big change is recognizable with the other two codebooks. Both systems trained with the use of transcriptions have the availability of the fricative states so that the estimation of the envelope results really close as confirmed by the reduced LSD in graphic in Fig. 7. The LSD is until 20 dB lower than the classic codebook in some points. At least 10 dB less are verified in the central part of the utterance.

As we anticipated before the dynamic of the set trained with transcriptions (and no HMM retouch, 2nd experiment) is quite slow. Even if asymptotically the estimation is faithful, late on-sets provoked by scarce reactiveness are cause of inefficient extension and off-sets can produce disturbing artifacts. The 2nd system (trained with transcriptions) has a strong tendency of conserving the fricatives, confirmed also by the listening experience. Fig. 8 shows that in the 2nd experiment the LSD increases at the beginning of the /s/, although it keeps lower than the other two natural training sets, symptom of latent dynamic. The 4th experiment instead clearly permits a large decrement of the distortion

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7423
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7423

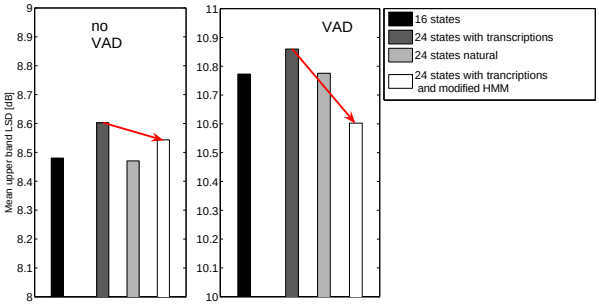

Figure 9: Mean LSD measures of the 4 experiments over the whole database, calculated over the whole speech (left) or over the active speech only (right). The arrow marks the difference in retouching the HMM matrix. Please notice the different LSD scales.

from the first frames. This is the result of the HMM thresholding which increases the percentage of a posteriori probability for the fricative states in Eq. 2 as soon as possible at the beginning of the utterance. A really helpful aggressiveness is found. The LSD remains about the same level as the rest of the speech in both 2nd and 4th experiment, which means no worsening in the extension of the fricatives, but in the latter one the faithful reproduction of the fricative envelope is exploited in the first two or three frames already. Regarding the peaks that we notice with naturally trained codebooks in correspondence to the fricatives, the gain is always beyond 10 dB.

During listening tests we can also notice a slight noise presence, due to the fact that the fricative orientation can provoke a slight unwanted rise of the envelope also for the other sounds. In any case neither the quality or the intelligibility are affected, whereas the syllables with fricative consonant are much more clear and understandable. Its is rather affected the average LSD.

The LSD gain locates in a little percentage of speech, so that finally no big improvement is verified for the 2nd and 4th experiments (Fig. 9) due to explanation given before. The worsening of the spectral distance does not produce disturb and worsening of the intelligibility and it is, in average, higher than the local improvement of the fricative. The HMM retouch does not change much the statistics.

Reducing the averaging range to the active speech only (with the use of a VAD, voice activity detector) the improvement is clearer. The speech pauses are not considered as useful understandable parts of the speech, so that a better statistic about the useful gain is given, in the 4th experiment, by the difference of 0.16 dB with the 1st one, and of 0.26 dB with the 2nd one.

If an /s/-phone classificator or a labeled database were available during the test system, the range of measure could be further reduced to the fricative sounds only and the LSD improvement could be more consistent with the scope of the fricative enhancement. It would be also more evident as it is from the listening experience and from the analysis in Figs. 7 and 8.

5. CONCLUSION

We carried out some experiments over the artificial bandwidth extension algorithm and revealed that the extension of the fricative sounds is quite ineffective. We discovered missed benefits and difficulties in recognizing the fricatives and therefore in estimating the correct upper band envelope.

We implemented a codebook design method that increases the statistical representation of high upper band energy envelopes by enlarging the codebook with high states dedicated to the fricatives. We trained the introduced states' statistics with an ad-hoc strategy that makes use of special training material and an HMM thresholding expedient for the the correct exploitation of the transitions.

The new trained codebook demonstrates a high fidelity in the expressiveness by appreciably reducing the LSD of the fricatives. It also shows a ready dynamic that reproduces the fricative characteristics coherently with the natural speech flow. LSD measures confirm the satisfying recovery of the loss quality as well as individual subjective listening tests.

6. ACKNOWLEDGMENTS

The authors would like to thank Patrick Bauer at the Technical University of Braunschweig (Institute for Communications Technology) for the coordinated work in support of our experiments and the processing of the training sets.

7. REFERENCES

[1] P. Bauer and T. Fingscheidt. An hmm-based artificial bandwidth extension evaluated by cross-language training and test. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4589–4592, March 31 - April 4 2008.

[2] P. Bauer and T. Fingscheidt. Speaker and language dependency of artificial bandwidth extension. In *Proceedings of the 34. Deutsche Jahrestagung für Akustik (DAGA 2008)*, 2008.

[3] J. A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, University of Berkeley, 1998.

[4] ITU-T, 2003. Rec. G.722.2, Wideband Coding of Speech at Around 16 kbits/s Using Adaptive Multi-Rate Wideband (AMR-WB).

[5] P. Jax and P. Vary. An upper bound on the quality of artificial bandwidth extension of narrowband speech signals. In *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, volume 1, pages I–237–I–240 vol.1, 2002.

[6] P. Jax and P. Vary. On artificial bandwidth extension of telephone speech. *Signal Processing*, 83(8):1707–1719, 2003.

[7] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve. The effect of highband harmonic structure in artificial expansion of telephone speech. In *Proceedings of Interspeech 2007*, pages 2497–2500, Antwerp, 2007. International Speech Communication Association. ISSN=1990-9772.

[8] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[9] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7423
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7423