COMBINED ESTIMATION/CODING OF HIGHBAND SPECTRAL ENVELOPES FOR SPEECH SPECTRUM EXPANSION

Yannis Agiomyrgiannakis and Yannis Stylianou

Institute of Computer Science
Foundation of Research & Technology Hellas
P.O.Box 1385 Heraklio, GR-711-10 GREECE
Heraklion, Crete
{jagiom, styliano}@ics.forth.gr

ABSTRACT

This paper addresses the problem of expanding the bandwidth of narrowband speech signals focusing on the estimation of highband spectral envelopes. It is well known that there is not enough mutual information between the two bands. We show that this is happening because narrowband spectral envelopes have an one-to-many relationship with highband spectral envelopes. A combined estimation/coding scheme for the missing spectral envelope is proposed, which employs this relationship to produce a high quality highband reconstruction, provided that there is an appropriate excitation. Subjective tests using the TIMIT database indicate that 134 bits/sec for highband spectral envelope are adequate for a DCR score of 4.41. This is an improvement of 22.8% over a typical estimation of highband envelopes using usual mapping functions, in terms of DCR score.

1. INTRODUCTION

In most of todays voice communication services, only the first 4 kHz of speech signal are carried, causing a natural limitation to transmitted speech quality. The bandwidth limitation mainly affects unvoiced parts of speech, i.e. consonants, which usually have an important energy distribution beyond 3 kHz. Subjective tests in [1] show the significance of the lost spectrum in the perceived quality of speech.

Ideally, an advanced voice terminal will be able to estimate the lost wideband spectrum from the transmitted one, without any infrastructure modification. Less ideally, both voice transmitting and receiving terminals will have to cooperate, by exchanging some information that describes the lost spectrum, at the cost of some extra bandwidth. In [2] the highband expansion required an additional overhead of 1.35-2.3 kbps. The combination of transmitting and estimating the lost spectrum will provide the necessary quality/bandwidth/complexity trade-offs of speech spectrum expansion algorithms.

The Speech Spectrum Expansion problem consists of two subproblems:

- modeling the lost highband speech.
- estimating the parameters of the model from the transmitted narrowband speech.

These two subproblems are interrelated, because it is up to the efficiency of the model to hide or to unveil the mistakes of the estimation process.

Most researchers use an AR source-filter model for the lost spectrum. In this paper the lost spectrum shall be referred as "highband". In [2] the highband excitation is formed by modulating white noise with the time envelope of 3-4 kHz transmitted speech signal. We verified that this is a very efficient way to produce a high quality excitation. A justification for such a method may be found in [3].

Many methods have been proposed for the estimation of the highband spectral envelope. Some of them make use of Vector Quantization (VQ) to map a narrowband codebook to a highband codebook [4], while other methods rely to statistical models [5], [6], [7] to predict the lost spectral envelope.

In [8] it is shown that there is only 0.6 to 0.8 bits of mutual information between narrowband and highband spectral envelopes, if the spectrum is parameterized with MFCCs (Mel Frequency Cepstrum Coefficients). On the other hand, it is well known that a parameterization of the spectral envelope leads to a lower bound of mutual information, as it is stated by the data processing inequality [9]:

$$I(X;Y) \geq I(S(X);T(Y)) \tag{1}$$

where X is the narrowband spectrum, Y the highband spectrum, $S(.)$ and $T(.)$ some parameterizations of these spectra. Therefore, there may be more mutual information under a different parameterization. However, in [10] many different highband envelope parameterizations have been tested, under several memoryless mapping methods, and neither succeeded more than a relatively minor improvement in terms of spectral distortion. Thus, even efficient memoryless X-Y mapping methods [11] are not enough for estimating the highband spectral envelope from the narrowband. Moreover, they provide little advantage over simple ones [4]. Therefore it seems that highband envelopes cannot be estimated from narrowband envelopes.

Under these conditions, two questions naturally arise:

- what is the reason for having so low mutual information?
- how much information is needed in order to have a high quality highband reconstruction?

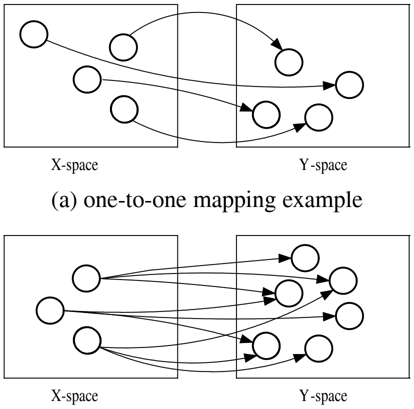
This paper addresses these two questions focusing only on an effective representation of highband envelopes. Section 2 shows that the relationship between narrowband spectral envelopes (X-space) and highband envelopes (Y-space) is one-to-many. In Section 3 we

This work was supported by the General Secretary of Research Technology Hellas and ICS FORTH, under ARISTEIA grand.

empirically quantify the extra information that is needed in order to have a high quality highband reconstruction. In Section 4 the one-to-many relationship is used in a combined estimation/coding scheme to encode highband envelopes with 4 bits/frame, at 33 frames/sec. Section 5 presents the highband expansion system used to subjectively evaluate our method. Finally, subjective tests in Section 6 show that the proposed method can achieve a 4.41 DCR score with only 134 bits/sec used for the highband envelope information.

2. THE RELATIONSHIP BETWEEN THE TWO BANDS

Highband and narrowband spectral envelopes have less than 1 bit of mutual information [8]. But, this is not enough for describing the highband envelope. The reason of this failure is that the basic hypothesis of estimation is badly violated in our case. This hypothesis is that one region of X-space maps to one region of Y-space. In that case, the estimation is one choice according to the pdf of the Y-space region, i.e. the expectation. In our case, even if the X-space is partitioned in very small regions, the corresponding Y-space region of every X-space region is rather large. One value -the "estimation"- cannot be representative of that region.



(a) one-to-one mapping example

(b) one-to-many mapping example

Fig. 1. Examples of one-to-one and one-to-many mappings between X and Y space

This can be shown schematically in Figure 1. In the case of one-to-one mapping (Figure 1a) between X-space and Y-space, a region in X-space maps to a region in Y-space. The size of the corresponding Y-space region defines the quality of the "estimation". The "estimation" is a representative of the Y-space region, and should be "optimal" under an appropriate criterion. This is the way memoryless estimation methods operate. A unimodal $P(Y|X)$ can enable a successful "estimation", provided that the variance of $P(Y|X)$ is small enough. A very large variance of $P(Y|X)$ implies multimodality, and a multimodal $P(Y|X)$ forbids good "estimations" if the classes in $P(Y|X)$ are distant. This is the case depicted in Figure 1b.

We argue that this is a usual mapping between narrowband and highband spectral envelopes. One way to visualize this is to perform a VQ on the Y-space region which corresponds to a particular X-space region. We have performed a VQ with 128 classes on X-space, randomly picked one class from these, and performed a VQ with 4 codewords to the Y-space envelopes which correspond to the X-space envelopes of that class. Figure 2 shows the narrowband envelope which represents the X-space class and the 4 corresponding highband envelopes. Picking another X-space class, and/or increasing the number of X-space classes, doesn't change the apparent variability of the highband envelopes which correspond to that X-space class. It is quite reasonable that an "estimation" in our case will neither capture spectral cavities or formants.

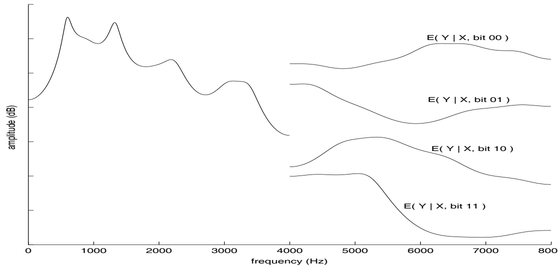


Fig. 2. X-Y space mapping example: a representative of a X-space class and it's 4 corresponding Y-space subclasses

3. QUANTIFYING THE NEED FOR INFORMATION

This section shows the effect of the one-to-many mapping in our estimation, and how this is related to the subjective quality of the reconstruction. Given a good excitation model that does not amplify errors in the envelope representation, we have found by informal listening tests that encoding the highband spectral envelope with 5 bits/frame is enough for a high quality highband reconstruction. This is consistent with the fact that higher frequencies are not as perceptually important as lower frequencies.

It is interesting to know how well mapping functions perform - in terms of distortion - and how far this performance is from having a good envelope estimation. Two memoryless mapping estimators were implemented; a simple VQ mapping called NLIVQ (Non Linear Interpolative Vector Quantization) [4] and the GMM Conversion Function estimator [11], [5]. An LSF (Line Spectrum Frequencies) representation of an AR (Auto Regressive) model was used for both narrowband and highband spectral envelopes. The NLIVQ method uses two equal-sized codebooks, one for narrowband LSFs and one for highband LSFs. The narrowband LSF vector is classified to the nearest narrowband codevector which in turn it is mapped to one highband codevector. The narrowband codebook is constructed by a binary split LBG VQ algorithm, and the highband codebook is constructed from the means of the highband vectors which correspond to each narrowband class vectors. The GMMCF estimator uses a GMM of the narrowband vectors to compute the least squares error solution to an experts-and-gates regression function. The estimator converts the narrowband LSFs to the highband LSFs, based on the assumption that each narrowband Gaussian maps to one highband Gaussian. The highband encoding is done with a variant of the well-known binary split LBG algorithm.

In Figure 3 we compare the encoding with the estimation of the highband envelope. The performance of the highband VQ and these two estimators is objectively evaluated using the Symmetric Kullback Leibler (SKL) distance, since this distortion measure reflects the perceptual differences between spectral envelopes [12].

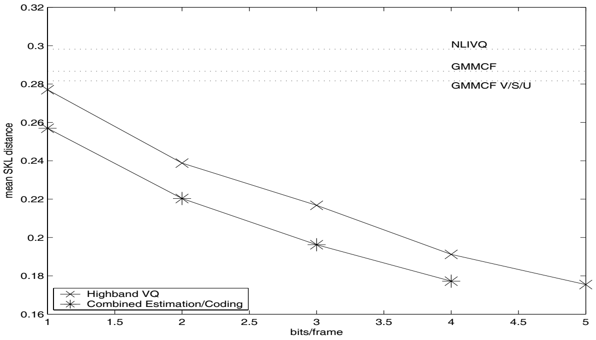


Fig. 3. Estimator Performance.

The SKL distance between AR spectra could be seen as a weighted formant distance [13]. The horizontal axis in Figure 3 reflects the number of bits used for encoding the highband, and the vertical axis represents the mean SKL distance. All the tests were carried out on the TIMIT database. For visualization reasons, the performance of the estimators (at 0 bits/frame) is displayed with horizontal dotted lines.

It can be clearly seen that the GMMCF estimator outperforms the NLIVQ estimator. Moreover the mean SKL distance obtained by using the GMMCF estimator is close to the distance obtained by using 1 bit/frame, which is consistent with the measured mutual information bound. If one GMMCF estimator is used for the voiced frames, another for the unvoiced frames, and a third estimator for the semivoiced frames, the distortion is reduced. But even with this additional voicing information the distortion measured at 1 bit/frame is not reached. Obviously this is far from having the high quality highband reconstruction observed at 5 bits/frame (mean SKL 0.18). The required distortion reduction is more or less equivalent to the transmission of 4 bits/frame.

4. COMBINED ESTIMATION/CODING SCHEME

The previous observations naturally lead to a combined estimation and coding scheme for retrieving the highband envelope. Initially we compute a VQ of the X-space. For each class in the X-space, we find the Y-space envelopes corresponding to the X-space envelopes belonging to that class. Then, these Y-space envelopes are vector quantized with K bits. The encoder, finds the nearest X-space class, classifies the Y-space envelope to one of the Y-space subclasses related to the selected X-class, and transmits the subclass index to the decoder. An example with a X-space class and it's 4 corresponding Y-space subclasses is depicted in Figure 4. In the receiver the narrowband spectral envelope from the transmitted signal is classified to the nearest X-space class, and a highband envelope is assigned according to the received subclass index of the corresponding Y-space subclasses. The VQ was computed with a variant of the standard binary split LBG algorithm.

Objective results of the proposed method, based on SKL distance, with 128 X-space classes and 1-4 bits/frame, are depicted in Figure 3 (lines with stems). The proposed method is efficient if it is taken into account that the mutual information is less than 1 bit, and that the knowledge of the narrowband envelope leads to a gain of approximately 1 bit over the performance of using just a VQ scheme for the highband.

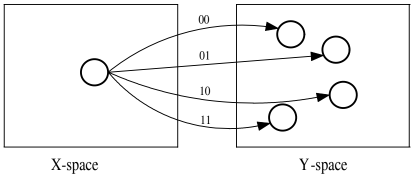


Fig. 4. Combined Estimation/Coding scheme for a 1-to-4 mapping.

5. HIGHBAND EXPANSION SYSTEM

A speech spectrum expansion system has been developed to test this combined estimation/coding scheme. The highband is modeled with a simple AR source-filter model. At the encoder, the wideband speech signal splits in two bands, the 0-4 kHz narrowband and the 4-8 kHz highband. Both bands are subsampled and subsequently filtered. The highband model consists of the energy ratio (in dB) between the narrowband and the highband signal, and 10 LSFs which parameterize the spectral envelope of the highband. The LSF parameterization is preferred for it's useful coding and stability properties. At the decoder, the wideband signal is resynthesized from the transmitted narrowband signal and the reconstructed highband, according to the energy ratio and the highband LSFs.

The highband signal is reconstructed by exciting the 10-th order AR filter with modulated white noise. The modulation is done with the time envelope of the 3-4 kHz transmitted narrowband signal. The synthesis is not done via OLA (OverLap Add), since OLA in the case of noise synthesis -our case- may introduce audible fluctuations [14], but with a variable lattice filter and sample by sample interpolation of the reflection coefficients.

If the highband envelope is excited with unmodulated white gaussian noise, the reconstructed wideband speech contains an unnatural noisy sound. Modulating with the time envelope removes this artificial sound and provides a high quality highband. The time envelope gives a pitch dependent temporal structure and thus a phase information to the white noise. The noise is better integrated when the noise bursts are concentrated around pitch closure instants [8].

The time envelope is computed by filtering the absolute value of the 3-4 kHz speech signal with a lowpass filter of 300 Hz cutoff frequency. When the highband spectral envelope is well estimated or coded, the modulation produces high quality wideband speech. To the contrary, highband envelope errors tend to be amplified, due to errors in the excitation signal. This is caused by rapid amplitude variations of the time envelope, mainly in unvoiced parts of speech. To cope with this, we follow a strategy similar to [2] and filter the time envelope with a lowpass variable filter controlled by a simple voicing criterion, based on the energy ratio.

We have subjectively evaluated the described highband expansion system for the three following cases:

- original highband LSFs.
- estimated highband LSFs by NLIVQ with 128 classes.
- estimated/coded highband LSFs by the proposed method with 134 bits/sec.

The first two subjective tests were conducted in order to determine an upper and lower bound, respectively, for the subjective quality of the proposed combined estimation/coding method. For the third test we used 128 X-space classes and 4 bits/frame (1-to-16 mapping). Since the evolution of highband envelopes in time is relatively slow, and the human ear is insensitive to errors in these higher frequencies, a frame rate of 33.3 frames/sec was found to be sufficient. Therefore the total bandwidth required is 134 bits/sec.

6. SUBJECTIVE TESTS

The evaluation was conducted with a DCR (Degradation Category Rating) test [15]. The subjects were presented with the *original wideband signal* and the reconstructed wideband signal, and were asked to vote the degradation of the latter according to the former.

All the tests were conducted with PHILLIPS SBC-HP800 headphones and a SoundBlaster Extigy sound card, in an office environment. Listeners were initially presented with written instructions and one example for each of the gradings: 5,3,1. Each listener received a different randomization of the stimuli. The initiation of each stimuli was made automatically, but the listener also had the option to reinitiate the stimuli by clicking a button. A short tone preceded the stimuli initiation to prepare the listener for the initiation. The listener voted by clicking a button, and a new stimuli was presented to him. All stimuli were energy normalized to the same acoustic level.

For the first two tests, 29 listeners participated, and they were asked to vote for 41 utterances from test set speakers; 14 utterances for the NLIVQ estimator, 14 utterances using the original LSFs, a null set of 5 stimuli used to check the bias of the listener [15], and 4 stimuli repeated for each test, to check if the listener had a consistent opinion. A few extreme cases of outlier listeners who obviously failed to the null set and to the repeated stimuli set, were excluded. The test of the combined estimation/coding scheme was conducted with 19 listeners, using 16 utterances from the test set, 4 repeated utterances and 5 null set utterances, under the very same conditions.

Table 1 states that using a synthetic excitation and the original LSFs produces a high quality wideband speech, almost indistinguishable from the original. The NLIVQ estimator did not perform well, as expected. The proposed scheme gets a very good DCR score which is close to the score obtained using the original LSFs. This shows that the highband envelope is well represented by only 134 bits/sec.

All the experiments in this paper were conducted using the TIMIT database training set (738431 samples) for training, and the TIMIT test set (271366 samples) for testing, preemphasis at the narrowband, 30ms windows, 14 LSFs for the X-space, and 10 LSFs for the Y-space.

Estimator	DCR score
NLIVQ estimator with 128 classes	3.59
Combined Est./Coding with 4 bits/frame	4.41
ORIGINAL highband envelope	4.67

Table 1. DCR test rating using the original wideband signal as reference

7. DISCUSSION

We gave an interpretation to the low mutual information between the narrowband and the highband spectral envelopes, by showing that each region in X-space is mapped to a number of regions in Y-space, which implies that the nature of $P(Y|X)$ is multimodal. This interpretation was used to build a combined estimation/coding scheme for the highband envelope which seems to use most of the available mutual information. Subjective tests indicate that with only 4 bits/frame, 134 bits/sec, a highband envelope estimation/coding of high quality (4.41 DCR score) is feasible. Sample utterances from this work may be found in:
http://www.ics.forth.gr/~jagiom

8. REFERENCES

[1] S. Voran., "Listener ratings of speech passbands," *IEEE Workshop on Speech Coding*, pp. 81–82, 1997.

[2] A. McCree, "A 14 kb/s wideband speech coder with a parametric highband model," in *Proc IEEE Int. Conf. Acoust.*, Istanbul, 2000, pp. 1153–1156.

[3] Jan Skoglund and Bastiaan Kleijn, "On time-frequency masking in voiced speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, July 2000.

[4] A. Gersho, "Optimal nonlinear interpolative vector quantization," *IEEE Trans On Comm.*, pp. 1285–1287, 1990.

[5] K.Y. Park and H.S. Kim., "Narrowband to wideband conversion of speech using gmm-based transformation," in *Proc. ICASSP*, Istanbul, June 2000.

[6] Peter Jax and Peter Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model," in *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong SAR, China, April 2003, vol. 1.

[7] Y.M. Cheng D. O'Shaughnessy and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, Oct. 1994.

[8] M. Nilsson, S.V. Andersen, and W.B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process*, Orlando USA, 2002.

[9] T. Cover and J. Thomas, *Elements of Information theory*, New York: Wiley, 1991.

[10] J. Epps, "Wideband extension of narrowband speech for enhancement and coding," *Phd thesis*, Sept. 2000.

[11] Y. Stylianou, O. Cappe, and Eric Moulines, "Continuous propabilistic transform for voice conversion," *IEEE Trans. Speech Audio Processing*, 1998.

[12] Stylianou Y. and Ann K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis.," in *Proc. ICASSP*, 2001.

[13] Raymond Veldhuis and Esther Klabbers, "On the computation of the kullback-leibler measure for spectral distances," in *IEEE Transactions on speech and audio processing*, Jan. 2003, vol. 11.

[14] Pierre Hanna and Myriam Desainte-Catherine, "Adapting the overlap-add method to the synthesis of noise," in *Proc. 5th Int. Conference on Digital Audio Effects (DAFx-02)*, Hamburg Germany, September 2002.

[15] K.K. Paliwal W.B. Kleijn, Ed., *Speech Coding and Synthesis*, Elsevier, 1995.