A WIDE-BAND SPEECH-MODEL PROCESS AS A TEST SIGNAL

M.R. Serafat and U. Heute

Institute for Network & System Theory, University Kiel, Germany

Tel: +49 431 77572 401, Fax: +49 431 77572 403, E-Mail: res@techfak.uni-kiel.d400.de

ABSTRACT

some of the major problems in objective quality assessment of speech coding systems or in testing other adaptive speech transmission systems are the speaker dependence, reproducibility, and the comparability of the measurement results, if natural speech is used as the test signal. This problem can be avoided by using suitable speech-model processes.

In this paper, we present a wide-band speech–model process, which includes the same long- and short-time characteristics as natural speech. The controlling part of the generator of this process involves several trained Markov chains (mc) to adapt the time-varying properties of the process to those of natural speech. Furthermore, special care is taken of the necessary probabilty density function (PDF) asymmetries, because the natural wide-band speech has an asymmetric PDF.

1. INTRODUCTION AND BACKGROUND

Up to now, the common methods for quality assessment of coded speech signals are based on highly standardized human listening tests. Because of several shortcomings (e.g. listener dependences) of this subjective method, many objective measures have been proposed and investigated in the past. Most of these approaches use natural speech as the test signal. Therefore, the results of these measures depend on the speech material used. To avoid this problem, a speech-model process is necessary, which includes the same characteristics as natural speech. Similar thoughts apply to test of other adaptive systems like, e.g., echo canceller.

This paper contains a brief description of our speech-model process MSMP (Markov Speech-Model Process) proposed as a test signal for wide-band (0–8kHz) speech-processing applications. This process is an extension of our known model for the narrow-band case (300–3400Hz), termed MSIRP (Markov Spherically Invariant Random Process)[1]. The MSIRP itself is an extension of Brehm's concept [2] to generate a narrow-band speech-model process. The contour lines of the bivariate probability-density function (PDF) of telephone speech are nearly of ellipsoidal or circular types [3]. This property was the motivation to model the telephone speech as a nearly spherically invariant random process (NSIRP). NSIRP included the same long-term properties as natural telephone speech, especially concerning its bivariate PDF. But this concept did not include any time-variant part. As an improvement of this concept, we proposed a modified NSIRP which also obtained some short-time properties of natural telephone speech like varying pitch and formant structures. Those modifications involved trained Markov chains to control the time-variant characteristics of the resulting process. But there were, on the one hand, some deviations between the long-time spectrum and the spectrum of the signal envelope of the MSIRP and those of natural telephone speech, and, on the other hand, MSIRP was limited to telephone-band signals. Threrefore, it could not be used for wide-band speech-processing applications: The concept of NSIRP and MSIRP used the special characteristics of narrow-band speech, which are different from those of wide-band speech, namely, a symmetric PDF and nearly spherical invariance. In order to obtain a process modelling wide-band speech appropriately, thorough modifications were necessary.

The modifications can be divided into 2 parts. The first part deals with the extension of the test signal for wide-band applications and the adaptation of its PDF to that of natural speech. The second part involves improvements of the long-term properties of our process; due to these improvements, the spectra of the signal and the signal envelope of MSMP now very closely approximate those of natural speech.

This paper is organized as follows: In section 2, a brief description of the generation procedure of MSMP will be given. In part 3, we present some measurement results and comparisons with natural wide-band speech. We conclude with some remarks in section 4.

2. GENERATION PROCEDURE OF MSMP

The generation procedure of MSMP is depicted in the block diagram of Figure 1. The MSMP is constructed as the product of a Gaussian process $\eta(t)$ and of a process $\sigma(t)$. This concept renders it possible to adjust the PDF, which is controlled in the upper branch, and the autocorrelation of the resulting process (controlled in the lower part) separately. The trained Markov chains form the main part of the controlling system. They are responsible for time-varying properties. The 'HMM-pitch' represents a hidden-Markov model (HMM), 'mc-formant' is a generalized mc, and 'HMM-energy' incorporates a HMM. Also 'mc-formant' works as the hidden part of 'HMM-energy'.

The decision, whether a frame of 30 ms duration is voiced (v) or unvoiced (uv), is carried out by the 'HMM-pitch', which is responsible for the pitch value in this frame. It has to be mentioned that an earlier version of MSMP involved only a simple Markov chain of first order to control the pitch variation. This mc was trained such that only the transition probability of the pitch values of MSMP was adapted to that of natural speech. But it did not consider the duration of the voiced or unvoiced regions and the relationship between them. Indeed, this earlier version of MSMP produced more switches between such regions than natural speech contains. To avoid this drawback, we substitute this mc with a HMM. The hidden part of this HMM is trained with respect to the duration of the different regions.

The formant filters are specified by a set of 100 lattice filters ($\mathrm{H}_1 \cdots \mathrm{H}_{100}$) of 16th order. The coefficients of these filters are obtained from a codebook, which was designed by an algorithm described in [4]. Depending on the decision of the 'HMM-pitch', 'mc-formant' chooses one of these filters and the 'HMM-energy' specifies a suitable gain-term. The pitch frequency can be varied by changing the parameters of the comb filter obtained from 'HMM-pitch'. This comb filter is described by $H(z) = \frac{\alpha}{1-a_k z^{-k_0}}$: The pitch frequency is determined by k_0, the sharpness of harmonically spaced peaks by a_k. The switch between voiced and unvoiced regions is carried out in two steps to achieve smoothed transitions: a_k equals 0.6 for the first and the last frame of each region and 0.95 in all other cases. For a smoothed switching between the formant filters, the actual filtering coefficents are updated every 2 ms by linear interpolation between the two coefficient sets of neighboring frames.

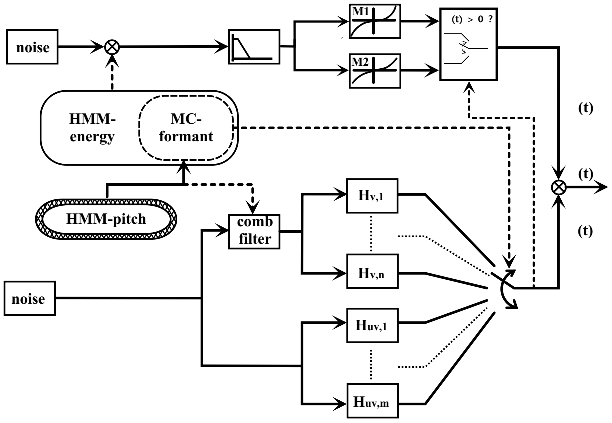

Figure 1: Block diagram of the MSMP generator

The output samples of the noise generator in the upper branch are multiplied with a suitably chosen gain-term which is obtained from 'HMM-energy'. As mentioned before, the value of the gain-terms depends on the decision of 'mc-formant', because this mc also represents the hidden part of 'HMM-energy'. Nevertheless we have also to smooth the switching of the gain-terms. Therefore, we also update the gain-terms every 2 ms by linear interpolation between the two gain-terms of neighboring frames. The resulting random process of the above multiplication is used as the input of the lowpass filter in the upper branch. This lowpass filter produces a slowly varying process to shape the signal envelope of the resulting process. Originally we needed a lowpass filter with the cutoff frequency of $f_s/3200$ (where f_s denotes the sampling frequency). But the design of a numerically stable lowpass filter with the above required frequency responce is very difficult. Furthermore, the impulse response of such a filter is very long. This causes some problems with the synchronization of both branches. Therefore, the generation of the desired output of this lowpass filter is carried out in two steps. At the first step, we produce a process with the sampling frequency of $f_0 = f_s/4$ and filter this process by a lowpass filter with the cutoff frequency of $f_0/800$. Thereafter, we convert the sampling rate to f_s. In this way, we reduce the numerical difficulties of the realization of the original filter. The other blocks in the upper branch form the process $\sigma(t)$ such that the multiplication of this process with $\eta(t)$ yields the desired PDF of natural speech. It is worth mentioning that special care has to be taken of the necessary PDF asymmetries, because natural wide-band speech has an asymmetric PDF. Therefore, we cannot approximate it with a product process in a simple way: *The PDFs of a product process are generally symmetric, if the participating processes are mutually independent and one of them has a symmetric PDF with zero mean.* Thus, we use two different nonlinear mappings in the upper branch und switch between them, depending on the sign of the process $\eta(t)$, to obtain the desired asymmetric PDF of the resulting process.

3. MEASUREMENT RESULTS AND COMPARISONS

The PDF, the long-term spectrum, and the spectrum of the signal envelope of natural speech are compared to corresponding characteristics of MSMP. All these measurement results have shown that the characteristics of MSMP agree well with those of natural speech. Fig. 2 shows that the spectrum of MSMP very closely approximates that of natural speech. The small deviation between the spectrum of the signal envelope of MSMP and that of natural speech at frequencies around 33 Hz is caused by switching the gain terms every 30 ms (Fig. 3). The comparison between the PDFs of the MSMP and natural speech is exhibited in Fig. 4. Obviously, natural speech has an asymmetric PDF, which is well reflected by the PDF of MSMP. In addition, we compared the synthetic sequence of pitch frequencies, gain terms, and formant indices, where each index specifies one formant filter, with the corresponding sequence of natural speech by using their PDFs and autocorrelations. Some results of these measurements are presented in Fig. 5-6. The results of these comparisons can be seen as a measure for the adaptation of the sequence behaviours of the MSMP to those of natural speech. Please note that unvoiced frames are indicated with the pitch frequency of 0 Hz.

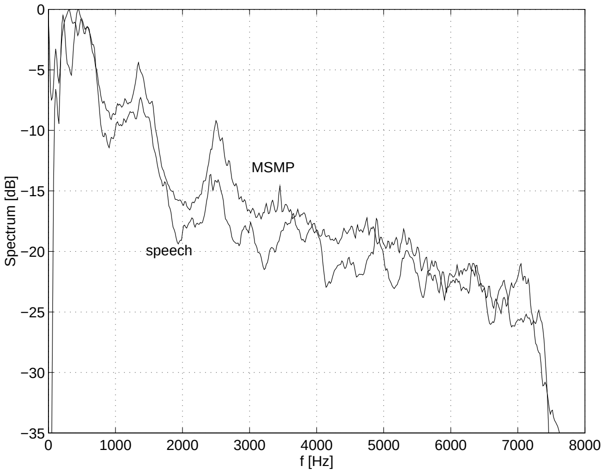

Figure 2: Spectrum of MSMP and natural speech

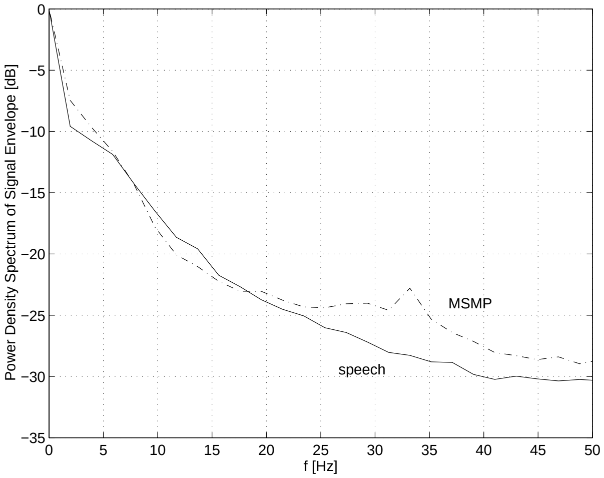

Figure 3: Spectrum of the signal envelope of MSMP and natural speech

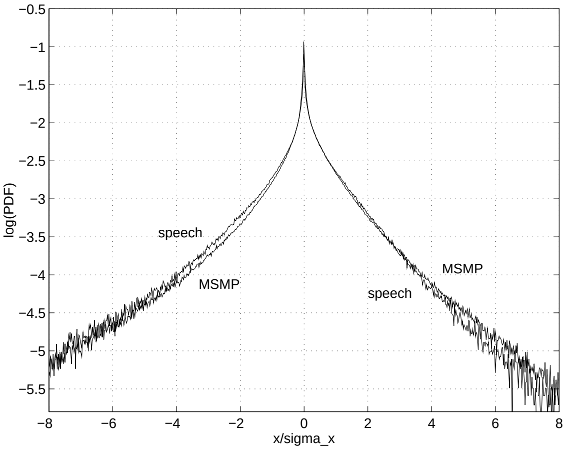

Figure 4: PDF of MSMP and natural speech

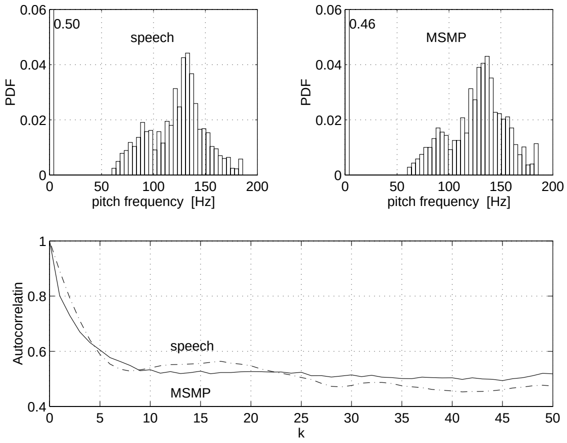

Figure 5: Autocorrelation and PDF of pitch values

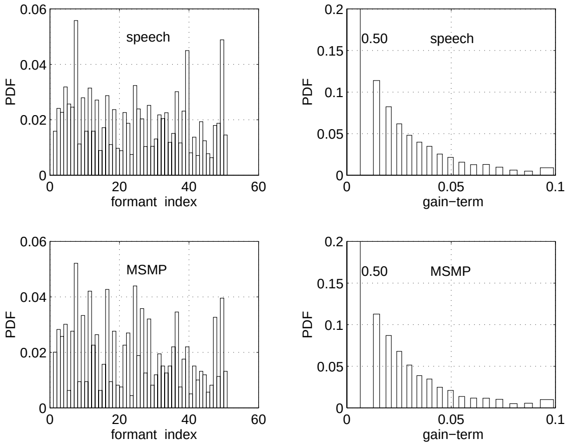

Figure 6: PDF of formant indices and gain–terms

4. CONCLUSION

The generation of a wide-band speech-model signal is shown to be feasible. MSMP exhibits typical natural-speech characteristics in good agreement. The above concept is applicable to produce a test signal which represents an 'average-speech signal'. It eliminates thereby, on the one hand, the dependences on the speech material in certain applications. On the other hand, the procedure renders it possible to include required special features if needed, like male/female-type speech, weak/strong variation of pitch frequency or even characteristics of one specific language. For such a purpose, we only have to use varying speech-data sets for the training of the Markov chains.

Furthermore, this concept allows us to express the univariate PDF of MSMP in form of a generalized Meijer's G-function. Our final aim is to introduce a complete mathematical handling of MSMP. This will be the subject of our further studies.

5. REFERENCES

[1] U. Halka and U. Heute, "Speech-Model Process Controlled by Discrete Markov Chains", Proc. Asilomar Conf. Sig. Syst. Comp., pp. 1196-1200, Pacific Grove, USA, 1993

[2] H. Brehm, W. Stammler, "Description and Generation of Sperically Invariant Speech-Model Signal", EURASIP, Signal Processing 12, Elsevier Science Publishers B.V., North Holland, pp. 119-141, 1987

[3] D. Wolf and H. Brehm, "Zweidimensionale Verteilungsdichten der Amplituden von Sprachsignalen", NTG-Fachtagungen, pp. 378-385, Erlangen, 1973 (in German)

[4] B. Juang, D.Y. Wong and A.H. Gray jr., "Distortion Performance of Vector Quantisation for LPC Voice Coding", IEEE Trans. Acoustic, Speech and Signal Processing, Vol. 30, No. 2, pp. 294-304, 1982