COMPARISON OF SPECTRAL MEASURE AND LISTENING TESTS RESULTS

Anna MADLOVÁ

Department of Radioelectronics, Faculty of Electrical Engineering & Information Technology
Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovak Republic

E-mail: madlova@elf.stuba.sk
Tel: (412-2) 60291743
Fax: (421-2) 65429683

Summary

The paper compares objective and subjective evaluation of synthetic speech quality. The RMS log spectral measure is used as an objective measure, and listening tests results are used as a subjective measure of synthetic speech quality. Four different methods of harmonic synthesis were used for evaluation: harmonic model with autoregressive or cepstral parametrization combined with simple concatenation of pitch-synchronous frames or overlap-and-adding the pairs of the same pitch-synchronous frames.

K e y w o r d s: spectral measure, spectral distortion, harmonic model, listening tests

Introduction

In speech processing the RMS log spectral measure is used to determine the error or difference between two spectral models on a log magnitude versus frequency scale [1]. Similar measure is the spectral distortion (SD), which has become the standard measure for evaluating the performace of spectrum coding. SD is defined as

$$\mathrm{SD}^2 = \frac{20^2}{f_2 - f_1} \int_{f_1}^{f_2} \left(\log_{10} \frac{\left| H\left(e^{j2\pi f / f_s}\right) \right|}{\left| \widetilde{H}\left(e^{j2\pi f / f_s}\right) \right|} \right)^2 df \qquad (1)$$

where f_s is the sampling frequency. $H(z)$ and $\widetilde{H}(z)$ are the original and quantized synthesis filters, respectively, for a current frame [2].

In this paper, instead of models or synthesis filters, the smoothed spectra of original and synthesized speech signals are used. Then, for $f_1 = 0$, $f_2 = f_s / 2$, and N_F-point FFT, the relation (1) may be rewritten for full-band spectral distortion as

$$\mathrm{SD} = \sqrt{\frac{2}{N_F} \sum_{k=0}^{N_F/2-1} \left(20 \cdot \log_{10} |P_k| - 20 \cdot \log_{10} \left| \widetilde{P}_k \right| \right)^2} \qquad (2)$$

where $\log|P_k|$ is obtained by smoothing $\log|S_k|$ of the speech spectrum given by

$$S_k = \sum_{n=0}^{N_F-1} w(n) \cdot s(n) \cdot \exp\left(- jn \frac{2\pi}{N_F} k \right) \qquad (3)$$

for $s(n)$ being samples of original speech, $\tilde{s}(n)$ being samples of synthesized speech, and $w(n)$ being samples of normalized weighting window. The weighted time-domain signal is usually zero-padded to increase frequency-domain spectral resolution. Different speech models can be compared by their spectral measure. The same speech models can be compared by a preference listening test of corresponding synthetic speech signals.

Methods

The speech material for spectral measure evaluation consisted of about 450 stationary parts of 5 vowels and 2 nasals spoken by male voice with the mean pitch frequency of about 110 Hz, sampled at 8 kHz. The spectral measure between smoothed spectra of original and resynthesized speech was computed using (2) from the speech frames weighed by a 24-ms Hamming window zero padded to 2048-point FFT given by (3).

Speech was resynthesized using the harmonic model with autoregressive (AR) or cepstral parametrization during analysis, and using concatenation of pitch-synchronous frames or overlap-and-adding (OLA) of consecutive pairs of the same pitch-synchronous frames during synthesis [3]. Combining these two methods of analysis and two methods of synthesis, four different synthetic speech signals were generated for each original speech signal. The speech material for listening tests evaluation consisted of twenty Slovak and Czech words spoken by the same male voice and synthesized by the same four combinations of methods. Words were grouped into pairs of the same word synthesized by two methods. The words of the pairs were grouped in the random order and listeners had to choose the word with better resemblance to the original. Eight independent listening tests were performed. Preferences of the methods are shown in Table 1.

couple of methods	AR concatenation vs. OLA	cepstral concatenation vs. OLA	concatenation AR vs. cepstral	OLA AR vs. cepstral
1st listener	2 : 18	1.5 : 18.5	6 : 14	9 : 11
2nd listener	2 : 18	3.5 : 16.5	2.5 : 17.5	10 : 10
3rd listener	5 : 15	9 : 11	5.5 : 14.5	13 : 7
4th listener	9 : 11	10 : 10	9.5 : 10.5	12.5 : 7.5
5th listener	6 : 14	11.5 : 8.5	10.5 : 9.5	14.5 : 5.5
6th listener	8 : 12	9.5 : 10.5	9.5 : 10.5	12 : 8
7th listener	12 : 8	9.5 : 10.5	11 : 9	11.5 : 8.5
8th listener	6.5 : 13.5	10.5 : 9.5	6.5 : 13.5	9 : 11
mean	6.07 : 13.93	7.93 : 12.07	7.36 : 12.64	11.36 : 8.64

Table 1 Preferences for the synthesis methods according to the listening tests of 20 words.

word	the best method according to the listener								highest score
	1st	2nd	3rd	4th	5th	6th	7th	8th	
aféra	AO	AO,CO	?	AC	AO	CO	AC	AC	AC,AO
fórum	CO	CC,CO	CC,CO	CC,CO	AO	AO	CC	CC	CC
gramo	CO	CO	AO,CO	AO	AO	AO	AO,CO	AO,CO	AO
kobra	AO,CO	AO,CO	?	?	?	CC	all	?	AO,CO
liga	AO	?	?	CO	CO	CO	?	?	CO
mařka	AO	AO,CO	AO	AC	CC	CC	AC	AC	AC,AO
naďa	CO	AO	AO,CO	AO	AO	AO	AC	AO	AO
náradiu	CO	AO,CO	?	?	?	?	?	?	CO
neguj	CO	CO	CC	AC,AO	?	?	AC,CO	CC,CO	CO
niekto	CO	AO,CO	?	AC,CC	AC	AC	AO,CO,CC	?	AC,CO
obor	AO	AO	AO	?	AO	AO	?	?	AO
očko	AO	AO	AO,CC	CC	AC	AC	?	?	AO
ovca	?	AO	AO	?	CC	CC	?	?	AO,CC
pauza	AO,CO	AO,CO,CC	AO,CO	CO	AO	AO	CC	CC	AO
racek	?	?	?	AC	AC	AC	?	AC.CC	AC
šifra	AC	CC	AC, CC	AO	AO	AO	?	AO	AO
spev	CO	AO	AO	AO,CC	CC	CC	all	AO,CO	AO
stĺp	AO,CO	CO	AO,CO	AC,CC	AC	AC	?	?	AC,CO
ufo	AO,CO	AO,CO	AO	AO	AC,AO	?	?	AO,CO	AO
ulietať	AO,CO	AO,CO	AO,CO,CC	CO	CO	CO	?	AO,CO,CC	CO
highest score	CO	AO	AO	AC,AO	AO	AO	AC	AO	AO

Table 2 Evaluation of the listening tests for every word through all the listeners, and for every listener through all the words.

The listening tests were also evaluated for every word through all the listeners, and for every listener through all the words. Results are shown in Table 2. Here, new abbreviations for the harmonic model were used:

AC = AR parametrization + concatenation
AO = AR parametrization + OLA
CC = cepstral parametrization + concatenation
CO = cepstral parametrization + OLA

The word “all” means that the listener regarded all four pairs of the same word as having the same quality. The question mark “?” means that there was some disagreement in the listener’s evaluation, e.g. AO and CO were regarded of the same quality, AC and CC were regarded of the same quality, AO was regarded as better than AC, but CC was regarded as better than CO. Table 2 shows that the highest score is given to the AR method with OLA.

analysis method		AR		cepstral	
synthesis method		concatenation	OLA	concatenation	OLA
spectral measure [dB]	mean	2.89	2.69	2.78	2.80
	std	0.99	0.64	0.73	0.63
score of listening tests [%]		16.79	31.61	25.71	25.89

Table 3 Spectral measure and listening tests results for four combinations of methods.

Results

The mean preferences of the listening tests were rescaled in such a way that the sum of listening tests scores for all four methods gives 100 %. Resulting comparison of spectral measure and listening tests scores is shown in Table 3. We can see that the highest score of listening tests corresponds to the lowest RMS log spectral measure and vice versa. The harmonic speech model with AR parametrization and OLA synthesis gives the lowest RMS log spectral measure and the highest listening tests score. The harmonic model with AR parametrization and concatenation of pitch-synchronous frames gives the highest RMS log spectral measure and the lowest listening tests score. The harmonic speech model with cepstral parametrization gives similar values for concatenation as well as OLA. All these results mean that use of the RMS log spectral measure is justified for determining perceptual resemblance of two speech signals.

Conclusion

Comparison of spectral measure and listening tests for harmonic speech model with AR and cepstral parametrization using concatenation as well as OLA of pitch-synchronous speech frames has shown that instead of rather exhausting listening tests consuming time of several independent listeners, the RMS log spectral measure or spectral distortion may be used as well. Subjective evaluation dependent on good ear can be replaced by quantitative comparison of the smoothed speech spectra.

References

[1] A. Gray, J. D. Markel: “Distance Measures for Speech Processing”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, no. 5, pp. 380-391, October 1976.

[2] J. Skoglund: From Modeling to Perception – Topics in Speech Coding, PhD Thesis. Chalmers University of Technology, Göteborg, Sweden, 1998.

[3] A. Madlová: Some Parametric Methods of Speech Processing, PhD Thesis. Slovak University of Technology, Bratislava, Slovakia, May 2001.