

# Perceptual Irrelevancy Removal in Narrowband Speech Coding

Marja Lähdekorpi<sup>1</sup>, Jani Nurminen<sup>2</sup>, Ari Heikkinen<sup>2</sup>, and Jukka Saarinen<sup>2,1</sup>

<sup>1</sup>Institute of Digital and Computer Systems,  
Tampere University of Technology, Tampere, Finland  
marja.lahdekorpi@tut.fi

<sup>2</sup>Speech and Audio Systems Laboratory,  
Nokia Research Center, Tampere, Finland  
{jani.k.nurminen, ari.p.heikkinen, jukka.p.saarinen}@nokia.com

## Abstract

A masking model originally designed for audio signals is applied to narrowband speech. The model is used to detect and remove the perceptually irrelevant simultaneously masked frequency components of a speech signal. Objective measurements have shown that the modified speech signal can be coded more efficiently than the original signal. Furthermore, it has been confirmed through perceptual evaluation that the removal of these frequency components does not cause significant degradation of the speech quality but rather, it has consistently improved the output quality of two standardized speech codecs. Thus, the proposed irrelevancy removal technique can be used at the front end of a speech coder to achieve enhanced coding efficiency.

## 1. Introduction

The exploitation of the psychoacoustic principles can ultimately lead a speech coding process into a perceptually optimal state in which the signal quality remains high despite a considerable reduction of the bit rate. Especially at low bit rates, it is very advantageous to avoid coding the perceptually irrelevant information. This work is aimed at processing narrowband speech signals in such a manner that the modified signal contains less perceptually unimportant information than the original, yet keeping the speech quality essentially unaltered. The determination of the irrelevant components is based on a psychoacoustic model for audio signals proposed by Johnston [1].

The basic idea behind the inherent existence of perceptual irrelevancy is the masking phenomenon that is present in all real-world auditory signals. Masking is a consequence of the finite frequency resolution of the human auditory system and, basically, it means the process by which the perception of one sound is suppressed by another, louder sound. The overall masking effect is mainly determined by the relative levels and frequencies of the maskee and the masker, as well as by their temporal characteristics. The nature of a sound also has a prominent impact on its masking capability. An approximate measure of the amount of masking can be obtained by evaluating a masking threshold. It indicates the sound pressure level at which a test sound is just audible in the presence of a masker. However, when considering complex signals such as

speech, the exact evaluation of the masking threshold is very difficult and coarse simplifications must be made.

Utilizing the results of psychoacoustic research in signal compression has been increasingly in focus during the recent years. However, it is not by any means a novel idea. Schroeder reported already in 1979 his method of exploiting the auditory masking effects in speech coders [2] and a part of his work is utilized by Johnston in the masking threshold calculation. Applications of masking models have been reported mainly from the field of audio coding, but some interesting experiments have also been made with speech. An Australian research group has improved the perceptual quality of coded speech, without increasing the bit rate, by incorporating simultaneous masking in the linear prediction [3, 4]. Examples of psychoacoustically assisted speech enhancement methods can be found in [5] and [6].

In this paper, a perceptual preprocessing method for narrowband speech signals is presented. Section 2 describes an implementation performing the proposed irrelevancy removal procedure. In Section 3, the method is evaluated using both objective measures and perceptual evaluation. Section 4 presents a possible application of the proposed method within a standardized code-excited linear prediction (CELP) speech codec and, in addition, results from a listening test with a mixed excitation linear prediction (MELP) codec. Section 5 provides a brief discussion on the preprocessing method and the test results. Finally, conclusions are drawn in Section 6.

## 2. Implementation

Originally, Johnston suggested his masking model to be used for audio signals sampled at 32 kHz. The data window length was 2048 samples (64 ms) with an overlap of 1/16th. Because the work presented here is aimed at processing narrowband speech, the model was adjusted to work sensibly with a sampling frequency of 8 kHz. Consequently, a frame length of 320 samples (40 ms) and an overlap of 50 % between the frames are used. Each input speech segment is multiplied with a Hamming window and transformed into its frequency domain representation via a fast Fourier transform (FFT). An FFT of 512 samples provides an adequate frequency resolution. A block diagram of the overall preprocessing method is shown in Figure 1.

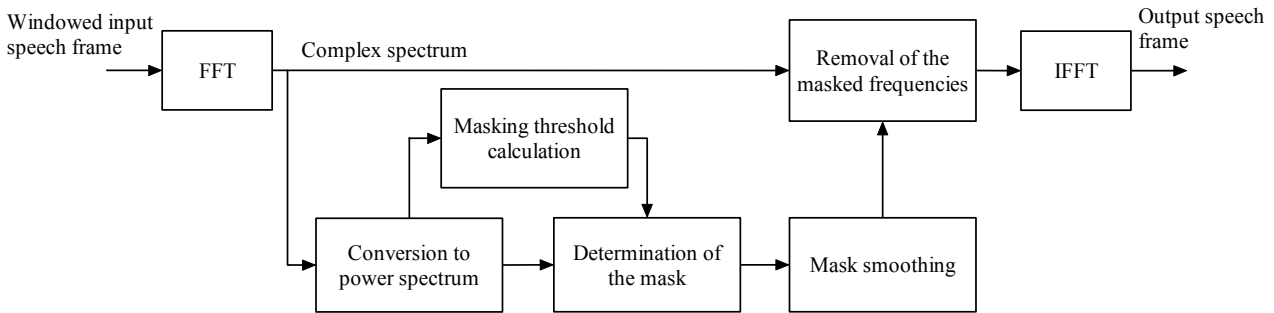


Figure 1. Block diagram of the preprocessing function.

The Johnston's model is used to determine the masking threshold for each segment of speech. The model provides a simple approach to the problem, taking only simultaneous masking into account. The power spectrum is divided into 18 critical bands according to [7] and a masking threshold is evaluated for each band. The final masking threshold is determined as the maximum of the calculated threshold and the absolute threshold of hearing at each critical band. An example of the thresholds for a female speech frame is shown in Figure 2. The masking thresholds of the frequency bands are compared with the power spectrum components to produce a binary mask. The mask is set to a value of zero at those frequencies where the power spectrum is below the masking threshold and a value of one is used elsewhere.

A straightforward means to remove the masked frequencies would be the multiplication of the complex spectrum of the input frame by the mask at each frequency. This corresponds to adaptive filtering of the input speech. However, due to the variation of the filter's frequency response (i.e. the mask) from frame to frame, the direct use of the mask would result in an output speech containing non-harmonic distortion caused by time domain aliasing [8]. Thereby, the mask is smoothed by convolving it with an appropriate window in the frequency domain. This procedure is detailed in [8]. The digital prolate spheroidal window [9] was chosen for this purpose because it is optimal in the sense that it concentrates most of its energy in the mainlobe and attenuates the aliasing components at its sidelobes, leading to a maximized ratio of the desired signal power to the distortion signal power. The design parameters of the digital prolate

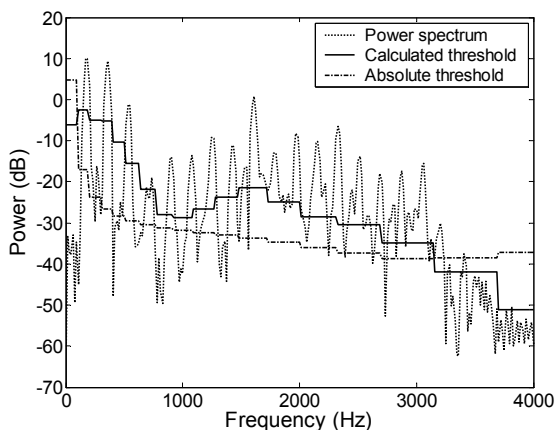


Figure 2. Example of the masking threshold determination.

window, the window length and the relative peak sidelobe height, were set to 7 samples and  $-39$  dB, respectively. This selection is quite prudent as the aim is to avoid excessive smearing of the mask.

After the mask smoothing, the output segment is adaptively filtered and transformed into time domain using the inverse Fourier transform. The overlapping sections are normalized so that the effect of the windowing is cancelled. The computational complexity of the proposed preprocessing technique is affordable for many applications: approximately 3.4 million floating point operations per second are needed, the FFT and its inverse transform being the most complicated parts.

### 3. Evaluation of the preprocessing method

The effects of the perceptual preprocessing on speech coding were first studied through some objective measurements. Due to the key position of the linear prediction (LP) in almost all modern speech coding systems, the LP analysis was performed on both the original and the preprocessed speech and the resulting LP residuals were examined. The LP coefficients were excluded from the tests since it has already been shown in [4] that the quantization of the LP coefficients calculated from the modified speech signal requires no extra bits compared to the LP coefficients calculated from the original speech. To verify that the preprocessing does not cause substantial deterioration of the speech quality, a listening test was performed using the comparison category rating (CCR) procedure [10]. The settings of the preprocessor were maintained as they were during the objective measurements in which an average of 43 % of the spectral components were deemed masked in each frame.

#### 3.1. Objective measurements

Using speech signals with the total duration of about 70 minutes, the energies and entropies of the different LP residuals were computed. The energy of the residual calculated from the preprocessed speech was, on average, 20.3 % smaller than that of the residual obtained by analyzing the original speech. The decrease in the energy causes modest loudness differences between the original and the preprocessed speech, but this effect can be compensated at the decoder using an additional level control.

By using the modified instead of the original speech in the LP analysis and filtering, the entropy of the residual signal was reduced by 7.2 %. This result together with the evident energy reduction confirms that the usage of the preprocessing technique in the front end of a speech coder produces a

residual signal that can be compressed more efficiently than the original residual signal.

### 3.2. Perceptual evaluation

Throughout the measurements described above, the preprocessing function was kept at constant settings so that the output speech did not suffer from substantial deterioration. This was verified through a CCR listening test. In the CCR test, listeners are presented with pairs of speech samples and for each pair, they are asked to grade the quality of the latter sample with respect to the former. The grades of the seven-point scale range from  $-3$  (much worse) to  $+3$  (much better). Each pair contains a processed sample and a quality reference that are presented in random order.

In this experiment, each sample was a single sentence. The test material consisted of Finnish speech filtered using an intermediate reference system (IRS) filter. Six female and six male speakers were chosen from a database and one sentence from each speaker was processed using the proposed technique. Furthermore, another version of each test sample was generated using an additional level adjustment in order to compensate for the slight loudness decrement caused by the preprocessing stage. Four reference sample pairs were also provided by choosing one male and one female sentence and processing them deliberately in such a manner that they had much lower quality than the original sentences. This was done with the preprocessing function using a random mask with 30–50% of the coefficients set to zero. Furthermore, in samples Male 1 and Female 1 (see Table 2) the smoothing window was not in use. Each processed sentence in the test had the corresponding direct connection version as the quality reference. The pairs were played in random order through high-quality headphones.

Altogether 24 naive listeners participated in the test. The listeners and the 24 actual test sample pairs were divided into three groups. The four reference pairs were common to all groups. Thus, each sample pair was listened by eight persons except for the reference pairs that were listened by all the listeners. In Table 1, the three listener groups are combined and the average grades for the processed samples with and without the additional level adjustment are shown. Table 2 presents the average scores given to the four references.

Table 1. Results of the CCR test. Average scores for the preprocessed speech with respect to the original  $\pm 95\%$  confidence interval.

Speaker gender	Without level adjustment	With level adjustment
Female	$-0.25 \pm 0.27$	$0.10 \pm 0.21$
Male	$-0.10 \pm 0.26$	$-0.08 \pm 0.26$
Total	$-0.18 \pm 0.18$	$0.01 \pm 0.17$

Table 2. Average scores of the reference samples.

Sample	Female 1	Female 2	Male 1	Male 2
Score	$-2.50$	$-2.50$	$-1.92$	$-1.75$

### 4. Application

In order to test how an actual speech codec performs with the preprocessed speech, an adaptive multi rate algebraic CELP codec [11] was used to code Finnish speech and segmental

Table 3. Segmental SNR (in dB) between the input and output of the codec.

kb/s	female			male		
	orig	prep	imp-%	orig	prep	imp-%
<b>4.75</b>	3.17	3.45	<b>8.71</b>	2.75	3.01	<b>9.54</b>
<b>5.15</b>	3.31	3.59	<b>8.56</b>	2.91	3.19	<b>9.55</b>
<b>5.9</b>	3.58	3.91	<b>9.25</b>	3.34	3.66	<b>9.78</b>
<b>6.7</b>	3.72	4.07	<b>9.64</b>	3.52	3.88	<b>10.11</b>
<b>7.4</b>	4.21	4.60	<b>9.21</b>	4.43	4.82	<b>8.80</b>
<b>7.95</b>	3.99	4.38	<b>9.72</b>	4.08	4.46	<b>9.37</b>
<b>10.2</b>	4.67	5.10	<b>9.23</b>	5.21	5.70	<b>9.47</b>
<b>12.2</b>	4.85	5.28	<b>8.76</b>	5.52	5.98	<b>8.43</b>

signal-to-noise ratio (SNR) was calculated between the input and the output of the codec. The test material consisted of about 39 minutes of female and 31 minutes of male speech. Both original and preprocessed speech were coded with each of the eight modes of the codec and the segmental SNRs were determined. Table 3 presents the SNR values and the improvement percentages when using the preprocessed speech instead of the original as the coder input. It can be clearly seen that the codec has performed better with preprocessed than with original speech even though no optimizations have been made to the codec that would support the handling of preprocessed speech.

To confirm the promising SNR figures, another listening test was arranged, this time using absolute category rating (ACR) [10]. In the ACR procedure, the listeners use a five-point scale to grade the quality of the samples that have been processed with the different test conditions. The average of all scores given to a particular condition yields the corresponding mean opinion score (MOS).

The test performed within this work contained 16 different conditions, including four modulated noise reference unit (MNRU) conditions with the noise level ranging from 8 to 26 dB. The remaining conditions tested the quality of the preprocessed speech and the performance of the CELP codec and a 2.4 kb/s MELP codec [12] with both original and preprocessed speech as the input signal. The test material was spoken by two male and two female speakers and two sentences were chosen from each. These eight samples were normalized to  $-26$  dBov and processed through each of the 16 conditions. Thus, each listener assessed the quality of 128 samples in random order. Altogether 14 naive listeners participated in this test. The MOS values for male and female speakers for all conditions are shown in Table 4 and the combined MOS values together with their 95% confidence intervals in Figure 3.

### 5. Discussion

The masking model often judges even more than 40% of the frequency components of a frame to be zeroed. Nevertheless, the results of the perceptual evaluation indicate that the preprocessing causes very little or no perceptual degradation. Even without the extra speech level amplification, the quality deterioration is hardly perceivable. When combined with the additional level adjustment, the speech quality remains essentially unaltered in the preprocessing, as can be seen from Table 1. The MOS difference between conditions 5 and 6 in the ACR test even implies a slight enhancement in the speech quality due to the preprocessing, but the magnitude of this particular difference in Table 4 should not be compared with

Table 4. Results of the ACR test.

Condition	Female	Male
01 MNRU Q = 8 dB	1.23	1.25
02 MNRU Q=14 dB	2.02	2.14
03 MNRU Q=20 dB	2.98	3.04
04 MNRU Q=26 dB	3.70	3.50
05 Direct	4.16	3.66
06 Preprocessed	4.41	4.09
07 CELP 5.15 kb/s original	3.54	3.32
08 CELP 5.15 kb/s preproc.	3.54	3.54
09 CELP 6.70 kb/s original	3.93	3.48
10 CELP 6.70 kb/s preproc.	4.21	3.73
11 CELP 7.95 kb/s original	4.13	3.79
12 CELP 7.95 kb/s preproc.	4.18	3.96
13 CELP 12.2 kb/s original	4.02	3.68
14 CELP 12.2 kb/s preproc.	4.27	3.96
15 MELP 2.4 kb/s original	2.48	2.64
16 MELP 2.4 kb/s preproc.	2.61	2.71

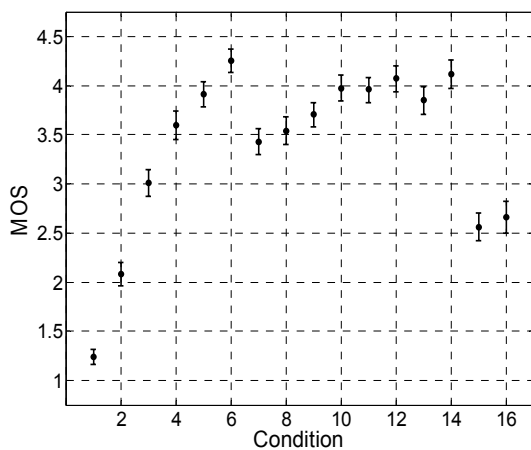


Figure 3. Combined MOS results with 95 % confidence intervals. The conditions are listed in Table 4.

the results of the earlier listening test because of the fundamental difference between the test methods.

The results of the ACR test indicate that the usage of the proposed preprocessing technique in the front end of a speech coder systematically improves the speech quality. Both waveform-approximating and parametric codecs have been tested and the direction of the change has remained the same. It should be noted that these results have been obtained without any modifications to the standardized speech codecs. Even better performance can be anticipated with appropriate optimizations.

## 6. Conclusions

In this work, a preprocessing method for narrowband speech signals was proposed. Utilizing a psychoacoustic model, the preprocessor determines the frequency components that are perceptually irrelevant due to the simultaneous masking. These components are then removed to reduce the amount of

perceptually irrelevant information in the speech signal. The effects of this procedure have been examined through both objective measures and perceptual evaluation, and the results have been promising. Firstly, the test results indicate that the usage of the proposed preprocessing method allows more efficient coding of the speech signal without significantly altering the speech quality. Secondly, the results have shown performance improvements in two different types of speech codecs when the preprocessor has been used as a front end to the coders.

## 7. References

- [1] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, Vol. 6, Iss. 2, pp. 314-323, February 1988.
- [2] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of the Acoustical Society of America*, Vol. 66, No. 6, pp. 1647-1652, December 1979.
- [3] J. Lukasiak, and I. S. Burnett, "Source enhanced linear prediction of speech incorporating simultaneously masked spectral weighting," *Journal of Telecommunications and Information Technology*, Vol. 2, pp. 15-23, December 2001.
- [4] J. Lukasiak, I. S. Burnett, and C. H. Ritz, "Low rate speech coding incorporating simultaneously masked spectrally weighted linear prediction," in *Proc. 2001 Eurospeech*, Aalborg, Denmark, September 2001, pp. 1989-1992.
- [5] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 6, pp. 497-514, November 1997.
- [6] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 2, pp. 126-137, March 1999.
- [7] E. Zwicker, and H. Fastl, *Psychoacoustics*, Springer-Verlag, Berlin, Germany, 1990.
- [8] U. Rass and G. H. Steeger, "Reducing time domain aliasing in adaptive overlap-add algorithms," in *Proc. 1999 138th Meeting of the Acoustical Society of America*, Columbus, Ohio, November 1999.
- [9] T. Verma, S. Bilbao, and T. H. Y. Meng, "The digital prolate spheroidal window," in *Proc. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, May 1996, pp. 1351-1354.
- [10] *Methods for subjective determination of transmission quality*, ITU-T Recommend. P.800, August 1996.
- [11] E. Ekudden et al., "The adaptive multi-rate speech coder," in *Proc. 1999 IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999, pp. 117-119.
- [12] L. M. Supplee, R. P. Cohn, and J. S. Collura, "MELP: The new federal standard at 2400 bps," in *Proc. 1997 IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, April 1997, pp. 1591-1594.