

FORWARD MASKING PHENOMENON IN CONCATENATIVE SPEECH SYNTHESIS

Miloš Cernák, Gregor Rozinaj

Faculty of Electrical Engineering
Slovak Technical University in Bratislava
(mcernak, gregor)@ktl.elf.stuba.sk

Abstract: *The approach described in the paper tries to get more knowledge to the concatenative text-to-speech system design. The knowledge is based on masking phenomenon of the inner ear, particularly of its temporal (forward) masking properties. Designing such knowledge-based system is suggested to use in the unit selection-based speech synthesis, as contemporary a prominent technique in concatenative synthesis, which utilizes a big speech corpus. The more prosodic variability the corpus captures, the more natural a synthetic voice sounds and there are more possibilities to occur a forward masking events during concatenation of selected candidate units from the corpus.*

Keywords: *Speech synthesis, human auditory masking*

1. INTRODUCTION

1.1 Human auditory perception

Incorporation of the results from psychoacoustics research into speech synthesis systems brings the knowledge about human's perception of speech, represented in acoustic level, to the synthesis process. There are well known two basic phenomena in speech perception:

- Frequency (simultaneous) masking
- Temporal (forward) masking

Frequency masking causes the low level signal, e.g. a small band noise (the maskee) can be made inaudible by simultaneously occurring stronger signal (the masker), e.g. a pure tone, if masker and maskee are close enough to each other in frequency [1]. This phenomenon is utilized in speech coding, where perceptually inaudible sounds needn't be coded, such as in MPEG-1 layer 1 standard. Tonal components are there identified as local maxima, which exceed neighboring components within a certain Bark distance by at least 7 dB [2]. Perceptually modified MFCC, as a new speech representation utilizes a frequency masking, was successfully used as a distance measure for costing spectral discontinuities in concatenative speech synthesis systems [3].

Temporal masking stands for dependency of perception of a current sound on preceding sound. If a loud sound is followed closely in time by weaker sound, the perceptibility of a weaker sound is diminished. Two time-domain phenomena play an important role in human auditory perception, pre-masking and post-masking. Whereas pre-masking tends to last about 5 ms, post-masking can last from 50 to 300 ms [2]. Temporal masking can be emulated by RelAtive SpecTrAl (RASTA) technique, as used in speech recognition [4].

1.2 A typical unit selection-based approach

Speech synthesizers can be classified either as rule-based systems or as data-driven systems. Concatenative synthesis belongs to data-driven systems, where speech feature vectors are derived from recorded speech taken from the speech corpus. The quality (i.g. the intelligibility and naturalness of synthetic speech) from the digital speech-processing point of view depends on:

- Chosen speech representation (e.g. lpc, mel-frequency cepstra, deltas, power coeff.)
- The way the representation is handled

The handling of the speech representation is usually made by defining cost functions [5], which are evaluated during synthesis process. The design of that cost functions should be perceptually motivated to model human auditory perception. This is a typical approach, based on short-term speech analysis, which describes the speech by a sequence of short-term feature vectors. Each individual feature vector is usually treated for cost function evaluation (every 10-20 ms) as independent of its neighbors. Temporal aspects of the speech signal are usually handled by defining context dependent (sub-) phoneme models.

We have found, there is small evidence in speech synthesis research that temporal aspects of speech, at least about syllable-length (around 200 ms), are confronted with human auditory perception and masking. In following section we are trying to identify those temporal aspects, at least forward masking phenomenon.

2. FORWARD MASKING

2.1 The problem of short-term analysis

Short-term speech analysis uses 10-20 ms frames. However, global features such as cepstral coefficients of spectral envelopes, every 10-20 ms, are easily affected by common frequency-localized random perturbation, which have hardly any effect on human speech communication [4]. The question is, what is beyond those 20 ms? As stated in introduction, post-masking can last from 50 to 300 ms. We have to know, where forward masking takes a place and eliminate it, additionally to the short-term analysis.

Forward masking effect (post-masking) is typically measured by presenting a masker (tone or band passed noise for 200 ms or longer), followed by a short signal called probe (a maskee), after variable delay [4]. The masking effect is measured by the rise of a threshold of detection of the probe.

2.2 Forward masking in concatenative synthesis

Unit selection-based speech synthesis usually uses a large acoustic inventory with rich prosodic (i.g. power, duration and F0) variations. Recorded speech is re-sequenced in accordance with input text and we claim, that the joining of candidate units can meet the conditions of an occurrence of forward masking effect. If forward masking resulted, the perceptibility of the second joined segment would be diminished (see Fig. 1). Then the temporal masking window there comes into existence.

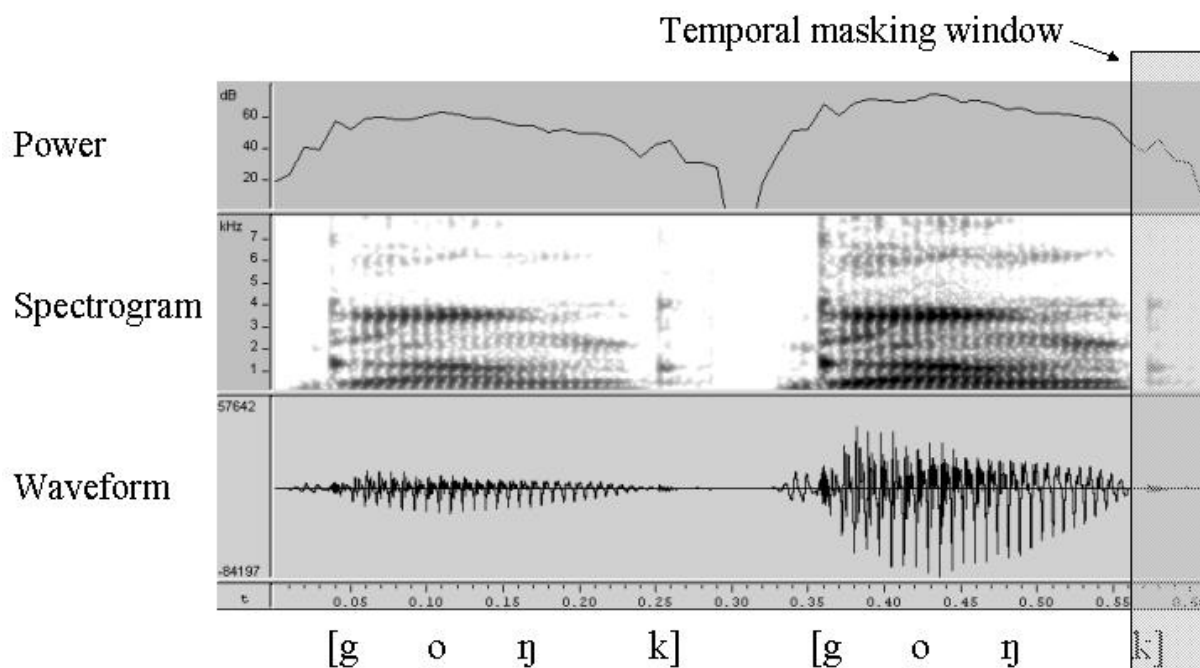


Fig. 1. Temporal masking window. There are shown the same words – gong, two times. The pronunciation is Slovak, with ‘k’ at the end. The concatenation of the segments ‘goŋ’ (200 ms) and ‘k’ (15 ms) on the left side is all right. The increased audio level at 10 dB of the segment ‘goŋ’ on the right side caused the creation of temporal masking window, and the perceptibility of ‘k’ was diminished.

From the example at Fig. 1, it seems that forward masking effect would occur, if the masker were a stronger signal, long enough. Such situation in unit selection-based synthesis can occur, because there can be more identically units with different power, even in the same context.

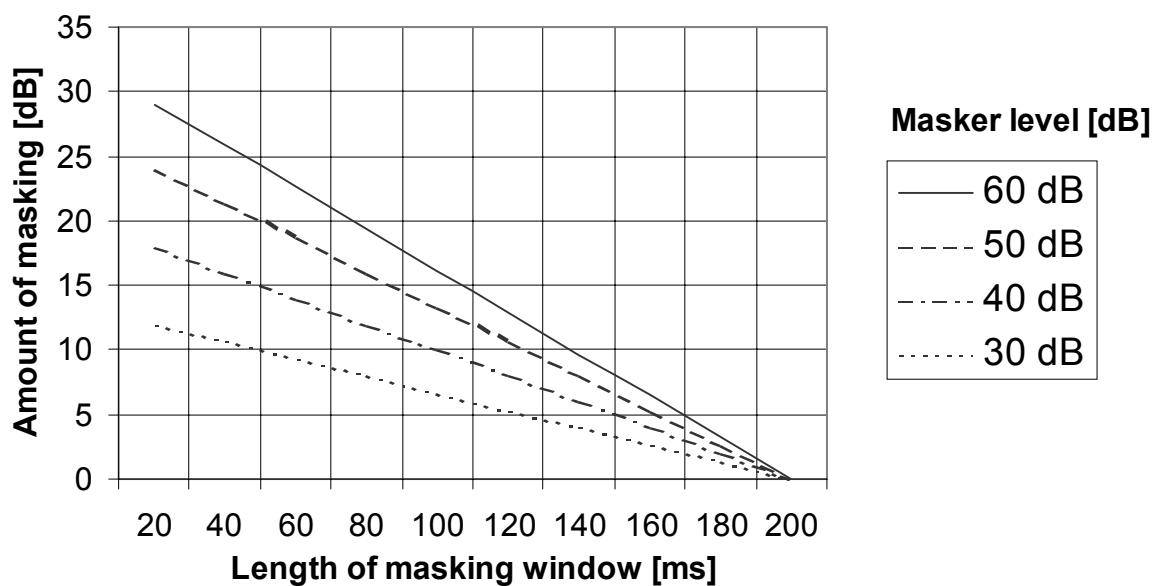


Fig. 2. Amount of masking and length of temporal masking window in terms of masker level

3. TEMPORAL MASKING WINDOW

For successive using of knowledge about forward masking in synthesis system, there has to exist some kind of ‘forward masking tracking’. In a word, we need to know the properties of temporal masking window. Its basic property is window’s length. Even though the window’s length from Fig. 1 is only about 15 ms, it masks the short plosive ‘k’.

The masker is identified as tonal components of speech, usually a syllable or a long vowel, which last about 200 ms. The power of adjacent candidate units is then checked and the decision about forward masking occurrence is made. The decision is based on predicted human performance, shown at Fig. 2. The results are taken from [4]. The properties of temporal masking window (length and the amount of masking of next sound) depend on the power level of masker.

4. CONCLUSION

The necessity of studying of temporal masking and other human auditory perception features in speech synthesis systems was claimed at NSF Workshop [6]. We rank our approach to knowledge-based technique, which tries to model of human auditory perception, not performing brute-force data mining in speech synthesis system.

Present synthesizers power normalize the speech database or include power specifications into the selection criterion; forward masking effect is minimized in this way. However, in this case real rich variability of the corpus is not fully utilized and we get more ‘neutral’ synthesized speech; the selection criterion of synthesizer doesn’t look 200 ms ahead.

In the paper we described forward masking phenomenon in concatenative speech synthesis system and we suggested how to utilize this knowledge during the synthesis process. Our objective was to point out, that also speech synthesizers ‘should have ears’. We see limitations of our current approach in searching and evaluating temporal masking windows; a more reliable algorithm should be designed. To emulate forward masking by means of RASTA technique is a promising way to do it. The confrontation of our approach with the definitions of the cost functions in unit selection synthesis ought to be made as well. The integration of using temporal masking windows into Slovak unit selection synthesis is our ongoing research.

REFERENCES

- [1] Haritaoglu, E.D., (2003) *Wideband Speech and Audio Coding*,
[<http://www.umiacs/umd.edu/users/desin/Speech/new.html>]
- [2] Huang, X., Acero, A., Hon, H.W., (2001) *Spoken Language Processing*, ISBN 0-13-022616-5, Prentice Hall PTR, New Jersey
- [3] Donovan, R.E., (2001) A New Distance Measure for Costing Spectral Discontinuities in Concatenative Speech Synthesizers, *4th ISCA Workshop, Scotland*
- [4] Hermansky, H., (1998) Should recognizers have ears? *Speech Communication* 25, 3-27
- [5] Hunt, A., and Black, A. (1996) Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database, *Proc. ICASSP’96, Atlanta*
- [6] Sproat, R., Ostendorf, M., Hunt, A., (1999) The Need for Increased Speech Synthesis Research, *NSF Speech Synthesis Workshop 1998*, The report