New Method for Delexicalization and its Application to Prosodic Tagging for Text-to-Speech Synthesis

Martti Vainio[1], *Antti Suni*[1], *Tuomo Raitio*[2],
Jani Nurminen[3], *Juhani Järvikivi*[4],*Paavo Alku*[2]

[1]Department of Speech Sciences, University of Helsinki, Helsinki, Finland
[2]Department of Signal Processing and Acoustics, Helsinki University of Technology
[3]Nokia Devices R&D, Tampere, Finland
[4]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
martti.vainio@helsinki.fi

Abstract

This paper describes a new flexible delexicalization method based on glottal excited parametric speech synthesis scheme. The system utilizes inverse filtered glottal flow and all-pole modelling of the vocal tract. The method provides a possibility to retain and manipulate all relevant prosodic features of any kind of speech. Most importantly, the features include voice quality, which has not been properly modeled in earlier delexicalization methods. The functionality of the new method was tested in a prosodic tagging experiment aimed at providing word prominence data for a text-to-speech synthesis system. The experiment confirmed the usefulness of the method and further corroborated earlier evidence that linguistic factors influence the perception of prosodic prominence.

Index Terms: prosody, delexicalization, speech synthesis, voice quality

1. Introduction

Delexicalization – removing segmentally relevant information from speech signals to render them unintelligible while retaining their prosodic characteristics – has a long history as a tool in prosody research [1, 2, 3]. The main reason for using delixicalized signals stems from the fact that listeners use lexical and grammatical information in making judgements concerning utterance internal phenomena which are related to prosody. There are a multitude of factors, such as for instance, word frequency, part of speech, as well as semantic content, which can influence the perceived prominence of words. Importantly, the perceived word prominences have been shown to be influenced by syntactic and information structure [4].

In Text-to-Speech synthesis (TTS), however, it is not realistic to model the very complex way in which human listeners perceive prosody when it is interacting with the linguistic content and grammatical structures of the utterances. Nevertheless, word prominence can serve as an intermediate parameter for a synthesis system [5] as the relationship between prosodic prominence and the acoustic parameters is more straight-forward.

One of the acoustic parameters which has been shown to be relevant for prosody in general is voice quality [6]. It has been fairly difficult to both analyze and model, but there are both subtle and large differences in voice quality that may be related to the perceived prominence of words: there are obvious spectral differences between stressed and unstressed syllables which are caused by different shapes of the glottal pulses as well as the fairly discrete modes (e.g., modal vs. non-modal voice). With respect to delexicalization, this calls for a method which leaves voice quality as well as the other prosodic parameters intact while removing the segmental information.

In this paper we present a new method for delexicalization of speech signals based on ongoing work in developing a new signal generation component for a Hidden Markov Model (HMM) based TTS system. The method uses inverse filtered glottal airflow pulses as an excitation for voiced speech retaining the prosodic and segmental characteristics related to the glottal pulse. After describing the method in detail, we present a pilot study in which we evaluated the functionality of the method by comparing prominence tags by experts to ones made by non-experts. The experiment aims at improving a procedure for obtaining training data for a TTS system that utilizes prominence for prosody control [7]. The system is trained using prominence annotations which are provided by experts and a subsequent statistical model to successfully tag a speech corpus for a HMM based synthesis system – tags provided by non-experts would greatly enhance the data preparation process in synthetic voice building.

2. Delexicalization using glottal flow based speech synthesis

The proposed delexicalization method is based on the parametrization and synthesis methods used in a recently developed HMM-based speech synthesizer [8, 9]. This speech synthesis system aims to produce high quality synthetic speech capable of conveying various styles of speaking, speaker characteristics, and emotions. To achieve this goal, the human speech production mechanism is modeled with the help of glottal inverse filtering. Glottal inverse filtering is a procedure where voiced speech signal is decomposed into the glottal source signal and the vocal tract filter. Through this decomposition, the behaviour of the natural glottal source signal can be modeled in the synthesis of the speech waveform.

The delexicalization procedure is illustrated in Figure 1. First, speech signal in parametrised into voice source and vocal tract features using the parametrization method of the TTS system. Then, the voice source is reconstructed from the parameters; natural glottal flow pulses are used in order to create the excitation signal, and this excitation is further modified in order to imitate the time-varying changes in the natural voice

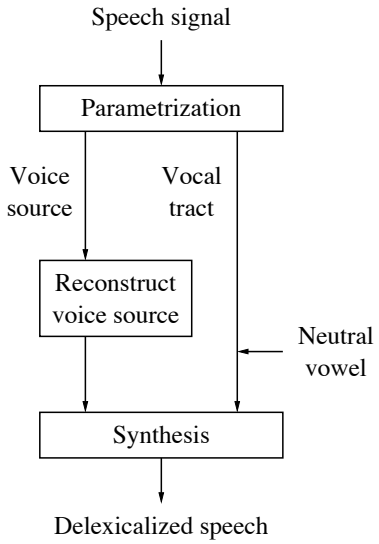

Figure 1: *Illustration of the delexicalization procedure. Speech is decomposed into voice source and vocal tract parameters. The voice source is reconstructed from the parameters and filtered with a modified vocal tract filter, where the varying parameters of voiced segments are replaced with constant parameters from a neutral vowel.*

source. Finally, speech is synthesized by filtering the excitation signal with the vocal tract filter, with the modification that the parameters of the varying vocal tract filter in voiced segments are replaced with constant parameters representing a neutral vowel. This procedure effectively delexicalizes speech but retains all the characteristics of the voice source, preserving all of the prosodic information including the voice source related changes. That is, in addition to fundamental frequency, features related to vocal effort such as voice quality, intensity and harmonic to noise ratio of the original speech are also preserved.

In the next few sections, glottal inverse filtering and speech parametrization and synthesis methods are explained in more detail.

2.1. Glottal inverse filtering

The basic idea of glottal inverse filtering is to separate the glottal source and the vocal tract filter based on the linear speech production model [10]. This theory assumes that the production of speech can be interpreted as a linear cascade of three processes: $S(z) = G(z)V(z)L(z)$, where $S(z)$ denotes speech, and $G(z)$, $V(z)$, and $L(z)$ denote the voice source, the vocal tract filter, and the lip radiation effect, respectively. Conceptually, glottal inverse filtering corresponds to solving the glottal volume velocity $G(z)$ according to $G(z) = S(z)1/V(z)1/L(z)$.

In this study, an automatic glottal inverse filtering method, Iterative Adaptive Inverse Filtering (IAIF) [11, 12] is used as a computational tool to implement glottal inverse filtering. IAIF is based on the repetitive procedure of canceling the effects of the vocal tract and the lip radiation from the speech signal. The only input required for the IAIF method is the acoustical speech signal recorded with a microphone. Various spectral modeling tools can be used within the IAIF method, but Linear Predictive Coding (LPC) is used in this work due to the computational efficiency and simplicity. The method is explained in more detail for example in [12].

2.2. Speech parametrization

The speech parametrization stage compresses the information of the speech signal into a few parameters which describe the essential characteristics of the original speech signal as accurately as possible. The voice source and the vocal tract are separately parametrized, enabling the individual modification of both speech production processes.

The parametrization is done as follows: First, the signal is high-pass filtered, and windowed with a rectangular window to 25-ms frames at 5-ms intervals. The speech features, presented in Table 1, are then extracted from each frame. The log-energy of the window is evaluated, after which glottal inverse filtering is performed with the Iterative Adaptive Inverse Filtering (IAIF) [11, 12] method in order to estimate the glottal volume velocity waveform from the speech pressure signal. IAIF iteratively cancels the effects of the vocal tract and the lip radiation from the speech signal using adaptive all-pole modeling. The outputs of the inverse filtering block are the estimated glottal flow signal $g(n)$ and the LPC model of the vocal tract $V(z)$. The spectral envelope of the glottal flow is parametrised with LPC (denoted by $G(z)$). The fundamental frequency is determined from the glottal flow signal with the autocorrelation method, and a harmonic-to-noise ratio (HNR) of four frequency bands (0–2 kHz, 2–4 kHz, 4–6 kHz, 6–8 kHz) is estimated from the glottal flow signal using cepstrum. HNR values are estimated through evaluating the cepstrum of each band, and comparing the energy of the cepstral peak, corresponding to the fundamental period, to the energy of other quefrencies of cepstrum. LPC models of the vocal tract $V(z)$ and the voice source $G(z)$ are converted to Line Spectral Frequencies (LSF) providing stability and low spectral distortion.

In case of unvoiced speech, conventional LPC is used to evaluate the spectral model of speech, and the speech parameters describing only voiced speech, namely, the fundamental frequency, harmonic-to-noise ratio, and the voice source spectrum, are not extracted.

2.3. Speech Synthesis

The flow chart of the synthesis stage is presented in Figure 2. The excitation signal consists of voiced and unvoiced sound sources. The basis of the voiced sound source is a glottal flow pulse extracted from a natural vowel. By interpolating the real glottal flow pulse according to F_0 and scaling in magnitude according to the energy measure, a pulse train comprising a series of individual glottal flow pulses with varying period lengths and energies is generated. For each pulse, the HNR is measured and the values are compared to the extracted HNR values. The amount of noise is matched in each frequency band by manipulating the phase and magnitude of the spectrum of each pulse. Furthermore, in order to mimic the natural variations in the voice source, the spectral tilt of each pulse is mod-

Table 1: Speech features and the number of parameters.

Feature	Parameters per frame
Fundamental frequency (F_0)	1
Energy	1
Harmonic-to-noise ratio (HNR)	4
Voice source spectrum $G(z)$	10
Vocal tract spectrum $V(z)$	20

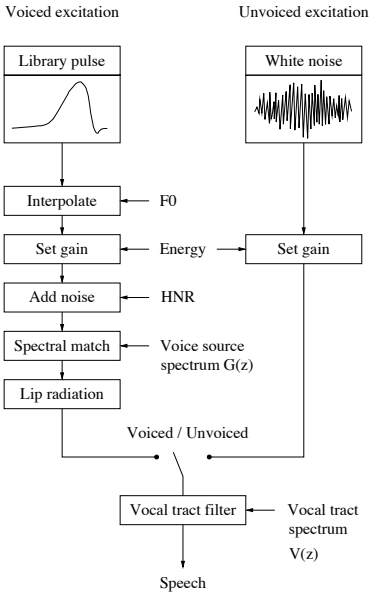

Figure 2: *Flow chart of the synthesis stage. The basis of the voiced excitation signal is a library glottal flow pulse, which is modified according to the voice source parameters. Unvoiced excitation is composed of white noise. Excitation signals are combined and filtered with the vocal tract filter $V(z)$ to generate speech.*

ified according to the all-pole spectrum. This is achieved by first evaluating the LPC spectrum of each pulse, and then filtering the pulse train with an adaptive IIR filter which flattens the spectrum of the pulse train and applies the desired spectrum. These procedures aim to preserve the original voice quality. For voiced excitation, the lip radiation effect is modeled as a first-order differentiation operation. Finally, LSFs are interpolated and converted to LPC coefficients $V(z)$, and used for filtering the excitation signal.

For delexicalization purposes, the LSFs of the vocal tract in voiced segments are replaced with parameters from e.g., a phonetically neutral vowel. The neutral vowel can be manually selected from the parameters, or it can be generated by evaluating the average of the vocal tract spectrum in voiced segments.

3. Word prominence labeling experiment

In order to evaluate the functionality of the new method we conducted an experiment, where a group of naïve listeners were asked to mark a set of words with their corresponding prominence. Such an experiment has not been conducted for Finnish, but e.g., Portele and Heuft ([5] and references therein) have reported fairly reliable agreements on prominence between naïve labelers. We were, therefore, interested in whether this would be the case with Finnish speakers, as well. More specifically, we were interested in whether the results would differ significantly between the original and delixicalized utterances with the working hypothesis that the results from the delexicalized ones would be – on average – closer to the labels provided by experts. In other words, if the method benefits the task of prominence tagging, we should observe less interference from non-phonetic linguistic information and closer agreement with expert opinion in the delexicalized condition.

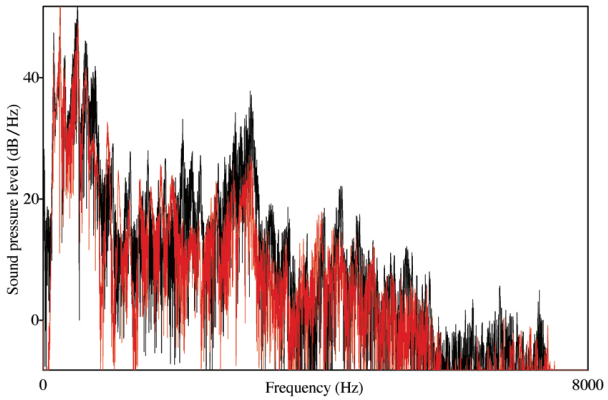

Figure 3: *Spectra from an original (red line) and a delexicalized (black line) utterances.*

Materials: Sixty utterances developed for a Finnish version of speech reception threshold test for speech coding evaluation [13] were used as test materials. The sentences were phonetically balanced and the recorded utterances were equated for intelligibility and, thus, had varying intensities. Materials from two speakers – one male, one female – were used in the current experiment.

Two experts (the first two authors) labeled the word prominences for each sentence; any disagreements between the experts were further discussed until an agreement was reached.

The chosen utterances were delexicalized using the new method in – what could be called – *semi-fricated* form. That is, unvoiced segments were replaced with low-pass filtered noise in order for the listeners to hear the ends of the utterances. Finnish speakers typically end their utterances with a non-modal voice (voiceless speech or creaky voice) after the last accented syllable (invariantly the first syllable of the last word). The constant vocal tract filter spectrum for a given utterance was obtained by first calculating the average spectrum of the syllable nuclei, and then selecting the LSF frame of the analyzed parameters that best matched the average spectrum. This was done in order to keep the spectral characteristics (in terms of e.g., slope) as similar as possible with the original signal.

The resulting delexicalized and the original utterances were divided into two counter-balanced sets consisting of 60 utterances each: 15 original and 15 delexicalized utterances from both speakers. The participants were randomly assigned to one of the two versions. Thus, none of the participants heard more than one version of a given experimental utterance.

Figure 3 shows spectra from the original (red line) and synthesized (black line) signals (515 ms from a voiced segment at the beginning of an utterance). The formant structure of the utterance has been almost totally disrupted while the overall characteristics of the original spectrum have been retained. Figure 4 shows a spectrogram of the a noun phrase “pieni näyttämö” (small scene) in its delixicalized form. The overall spectral shape of the original speech is preserved whereas the formants are kept stationary. The differences in harmonic-to-noise ratio are clearly visible between the different parts of the utterance.

Participants and procedure: 18 female students of phonetics, language technology, and logopedics took part in the experiment (ages 20-48 with an average of 26.9 years). All participants were native speakers of Finnish, none reported any hear-

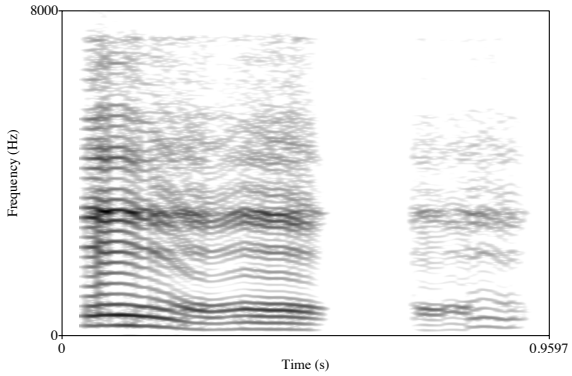

Figure 4: *A narrow-band spectrogram of the word pair "pieni näyttämö" (small scene). See text for more detail.*

ing problems, and none were acquainted with prosody related research. They could, therefore, be considered as naïve listeners.

The materials were presented to the participants using the Praat program and high quality headphones (Sennheiser HD-250). The participants could adjust the output volume of the headphones. The utterances were segmented on word level and the participants were instructed to listen to both the individual words and the whole utterance. No time limit was given and the test took between 30 to 60 minutes to complete. The participants were further instructed to mark on paper each words sentence stress on a scale from 0 (totally unstressed, typically e.g., a conjunction or a copula) to 3 (emphatically stressed; typically a narrow contrastive focus).

Results: Before the statistical analyses we had to discard the data from two speakers, who had systematically marked a wrong number of words in the sheet. There were an additional 24 missing items. This left 3621 responses for the analyses. The results were calculated as the absolute difference between the given prominence value and the one given by the experts. Table 2 shows the average absolute error for both delexicalized and original utterances for the final, initial, and medial word positions. In all cases the average error was fairly small and stays within one category. Obviously, the listeners were able to fulfill the task with a high degree of agreement. Importantly, the error in the delexicalized condition was smaller in all positions with the words in the initial position labeled most accurately.

Table 2: Average absolute errors in final, initial, and medial word positions for delexicalized and original utterances.

	final	initial	medial
Delexicalized	0.5659955	0.4854586	0.5475410
Original	0.7299107	0.5736607	0.6168122

We used a non-parametric Wilcoxon rank sum test to assess the significance between the *Original* and *Delexicalized* conditions. As expected, the difference is highly significant ($W = 1502188, p < 0.0001$). Thus, we can safely conclude that the naïve listeners behaved more like the experts when they were not influenced by the linguistic properties of the utterances.

4. Conclusions

In this paper we have described a new method for delexicalization of speech signals based on parametrization and synthesis methods that utilize a realistic voice source model. The new method allows for an unprecedented degree of control of both the glottal and vocal tract parameters in the process of delexicalization. We have furthermore shown in a listening experiment that the method can be successfully used in studying labelling of word prominence in running speech by naïve listeners. Therefore, the method improves the labeling accuracy to a degree that the labeling of large speech corpora for both general prosody related and TTS research could be partly done by phonetically untrained, non-professional, personnel. Importantly, the result provides further evidence that abstract linguistic properties have a strong influence on the perceived prosodic prominence of the individual words in an utterance.

5. Acknowledgements

The present study was supported by grants no. 107606, 125940, and 128204 from the Academy of Finland to the first author.

6. References

[1] I. Lehiste and W. Wang, "Perception of sentence boundaries with and without semantic information," *Phonologica*, vol. 19, pp. 277–283, 1976.

[2] J. Ohala and J. Gilbert, "Listeners'ability to identify languages by their prosody," *Report of the Phonology Laboratory Berkeley, Cal.*, no. 2, pp. 126–132, 1978.

[3] G. Sonntag and T. Portele, "PURR- a method for prosody evaluation and investigation," *COMPUT SPEECH LANG*, vol. 12, no. 4, pp. 437–451, 1998.

[4] M. Vainio and J. Järvikivi, "Tonal features, intensity, and word order in the perception of prominence," *Journal of Phonetics*, vol. 34, pp. 319 – 342, 2006.

[5] T. Portele and B. Heuft, "Towards a prominence-based synthesis system," *Speech Communication*, vol. 21, no. 1-2, pp. 61–72, 1997.

[6] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *15 th International Congress of Phonetic Sciences*, 2003, pp. 2417–2420.

[7] M. Vainio, A. Suni, and P. Sirjola, "Accent and prominence in Finnish speech synthesis," in *Proceedings of the 10th International Conference on Speech and Computer (Specom 2005)*, G. Kokkinakis, N. Fakotakis, E. Dermatos, and R. Potapova, Eds. University of Patras, Greece, October 2005, pp. 309–312.

[8] T. Raitio, "Hidden Markov model based Finnish text-to-speech system utilizing glottal inverse filtering," Master's thesis, Helsinki University of Technology, 2008.

[9] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, 2008.

[10] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed. Springer-Verlag, 1972, vol. 1.

[11] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[12] P. Alku, H. Tiitinen, and R. Näätänen, "A method for generating natural-sounding speech stimuli for cognitive brain research," *Clinical Neurophysiology*, vol. 110, pp. 1329–1333, 1999.

[13] M. Vainio, A. Suni, H. Järveläinen, J. Järvikivi, and V.-V. Mattila, "Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish," *Journal of the Acoustical Society of America*, vol. 118, no. 3, 2005.