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

Martti Vainio¹, Antti Suni¹, Tuomo Raitio², Jani Nurminen³, Juhani Järvičivi⁴, Paavo Alku²

¹Department of Speech Sciences, University of Helsinki, Helsinki, Finland
²Department of Signal Processing and Acoustics, Helsinki University of Technology
³Nokia Devices R&D, Tampere, Finland
⁴Max 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 delexicalized 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 straightforward.

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 parametrized 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.
Delexicalized speech can be interpreted as a linear cascade of three processes: the voice source, the vocal tract, and the lip radiation effect, respectively. Conceptually, glottal inverse filtering is to separate the glottal flow pulse according to the energy measure, a pulse train comprising a series of individual glottal flow pulses with varying period lengths according 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 modified.

<table>
<thead>
<tr>
<th>Table 1: Speech features and the number of parameters.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Fundamental frequency ( F_0 )</td>
</tr>
<tr>
<td>Energy</td>
</tr>
<tr>
<td>Harmonic-to-noise ratio (HNR)</td>
</tr>
<tr>
<td>Voice source spectrum ( G(z) )</td>
</tr>
<tr>
<td>Vocal tract spectrum ( V(z) )</td>
</tr>
</tbody>
</table>
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 noun phrase “pieni näyttäjä” (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-
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.

<table>
<thead>
<tr>
<th></th>
<th>final</th>
<th>initial</th>
<th>medial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delexical</td>
<td>0.5659955</td>
<td>0.4854586</td>
<td>0.5475410</td>
</tr>
<tr>
<td>Original</td>
<td>0.7299107</td>
<td>0.5736607</td>
<td>0.6168122</td>
</tr>
</tbody>
</table>

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