Frequency-Domain Delexicalization using Surrogate Vowels

Alexander Kain and Jan van Santen

Center for Spoken Language Understanding
Oregon Health & Science University, Portland, Oregon, USA
kain@cslu.ogi.edu, vansanten@cslu.ogi.edu

Abstract

We propose a delexicalization algorithm that renders the lexical content of an utterance unintelligible, while preserving important acoustic prosodic cues, as well as naturalness and speaker identity. This is achieved by replacing voiced regions by spectral slices from a surrogate vowel, and by averaging the magnitude spectrum during unvoiced regions. Perceptual tests were carried out comparing sentences that were either unprocessed or delexicalized, using a baseline or the proposed method. An intelligibility test resulted in a keyword recall rate of 92% for the unprocessed sentences, and near complete unintelligibility for both delexicalization methods. Affect recognition was at 65% for unprocessed sentences, and 46% and 49% for the baseline and the proposed method, respectively. Preference tests showed that the proposed method preserved drastically more speaker identity, and sounded more natural than the baseline.
Index Terms: delexicalization, intelligibility, affect

1. Introduction

The goal of *delexicalization* is to render the lexical content of an utterance unintelligible. This is achieved by removing from the speech signal the segmental features needed to recognize phonemes (and hence words), while preserving supra-segmental features such as pitch. A common application of delexicalization is to enable listening experiments on prosody in which word content is eliminated as a potentially confounding factor.

While it is easy enough to render an utterance unintelligible, preservation of the supra-segmental features is challenging because these features include more than just fundamental frequency (F0), duration, and energy. It is well-known that features such as spectral tilt, spectral balance, the degree to which vowel formants reach their targets, and various aspects of voice quality (e. g. tenseness), convey important prosodic information, such as word emphasis, and, more broadly, pragmatic prosody and affective prosody. Yet, typical delexicalization methods such as band-pass filtering eliminate most of these features. This is not a problem when the object of a study is that of, for example, verifying if a Text-to-Speech (TTS) system emphasizes the correct syllables.

However, in the study that motivates the methods discussed in this paper, far more acoustic prosodic features need to be preserved. Specifically, our study focuses on whether in individuals with autism there is a "disconnect" between lexical and prosodic content. This will be assessed by one group of judges labeling affect of speech transcriptions, and by another group of judges labeling affect of delexicalized forms of the same speech separately, and then applying temporal correlation methods to the thus produced label streams. To make this possible, we need a delexicalization method that preserves far more prosodic features, as affect is not conveyed purely through F0, duration, and energy, but also through voice quality, and many other additional, including currently not fully understood, features.

In this work, we propose an algorithm that represents a first step towards a process that, in addition to preserving more of these prosodic features, also preserves short-term acoustic features important to naturalness and speaker identification.

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders, 1R21DC010239 (Lois Black, PI) and from the National Science Foundation, 0905095 (Jan van Santen, PI). The views herein are those of the authors and do not reflect the views of the funding agencies.

2. Previous Approaches

The simplest delexicalization algorithms used pulse trains (e. g. impulse, sawtooth, sinusoid with harmonics, or glottal pulses) with F0 and amplitude identical to the original speech signal (with pauses during unvoiced sections) [e. g. 1, 2, 3]. Similarly, Maidment [4] used an electroglottograph signal, which represents the variations in glottal electrical resistance, and is therefore closely related to the original glottal waveform, uninfluenced by the vocal tract resonance or supra-glottal noise sources.

Lehiste and Wang [5] used a spectral inversion technique [6], which involves changing the sign of every other sample, thus inverting the spectrum without altering F0. The resulting speech sounds highly unnatural [7]. Lehiste [8] and Schaffer [9] used a bandpass filter that transmitted frequencies in the expected F0 range (e. g. between 70–270 Hz for a male speaker). However, segmental contrast is still perceived, such as vowel durations. Kreimann [7] used both spectral inversion and low-passed filtering. First the signal, sampled at 8 KHz, was high-pass filtered at 600 Hz, then inverted, then low-pass filtered at 2000 Hz. This resulted in a reportedly very speech-like, but quite incomprehensible signal. Sonntag and Portele [10] reviewed these algorithms and found that listener performance was similar across all delexicalization algorithms in their task. However, a subjective evaluation led the authors to conclude that F0-driven sinusoid including the first two harmonics (at 1/4 and 1/16 amplitudes, respectively) performed the best in terms of ease and naturalness of listening.

In contrast to approaches that operate on the signal directly, it is also possible to modify acoustic features in an analysis-synthesis framework: Nicolas and Roméas [11] used a a formant synthesizer to produce a steady-amplitude /ɑ/ with original F0 values for voiced regions, with 30 ms amplitude transitions between voiced and silenced unvoiced regions. Similarly, de Pijper and Sandermann [12] used a 16^{th}-order LP filter whose peaks were fixed at $500, 1500, 2500, \ldots, 7500$ Hz, with

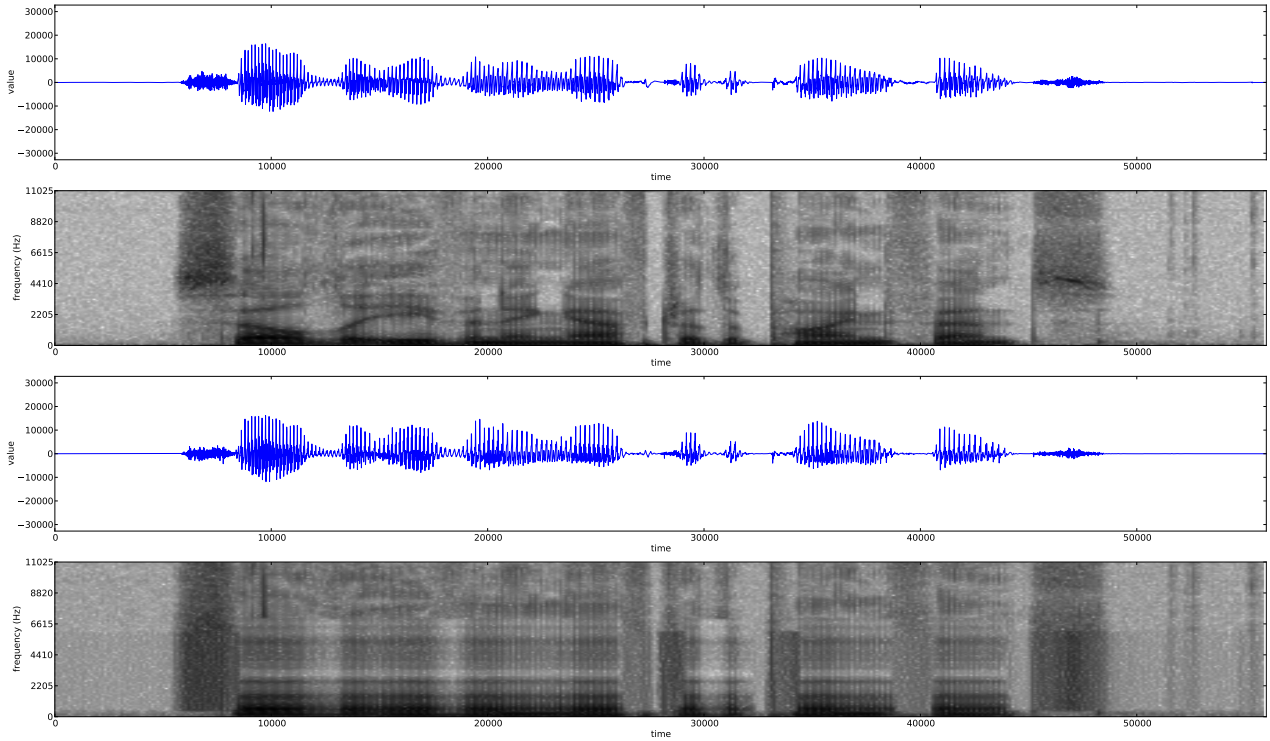

Figure 1: Waveforms and spectrograms of original (top panes) and delexicalized (bottom panes) speech.

bandwidths fixed at 50, 150, 250,. . ., 750 Hz. The energy and pitch trajectories (manually corrected) were extracted from an original speech waveform. Resynthesis resulted in reportedly high quality, unintelligible speech.

The disadvantages of the above approaches are that they destroy certain important prosodic features, such as spectral balance. To address this, Vainio et al. [13] recently described a system which incorporates inverse filtered glottal flow and all-pole, linear prediction (LP) modeling of the vocal tract, thus allowing a large degree of sensitivity and control. Specifically, they parametrized a voiced frame of speech into the following parameters: F0, log-energy, harmonic-to-noise ratio (HNR) of four frequency bands (0–2, 2–4, 4–6, 6–8 kHz), voice source spectrum (modeled as 10^{th}-order LP filter), and vocal tract spectrum (modeled as 20^{th}-order LP filter), using an iterative adaptive inverse filtering method. Unvoiced frames were modelled using a single LP filter. During synthesis, the excitation consisted of either scaled white noise, or a voiced sound source, constructed as follows: starting with a glottal flow pulse, the signal was interpolated according to a specified F0, and scaled according to a specified energy. Then, the amount of noise was matched in each frequency band in accordance with the desired HNR. Furthermore, spectral tilt was modified in accordance with a specified voice source spectrum. Finally, lip radiation effects were applied. Both voiced and unvoiced excitation sources were then filtered by the specified vocal tract filter, whose parameters are those of a phonetically neutral vowel. In our experience, the disadvantage of this approach lies in the relatively error-prone glottal flow estimation for new speakers, and the fact that the resulting speech naturalness is degraded (e. g. consonants are unnatural, vowels sound pressed).

Finally, it is also possible to carry out the transformation on a symbolic level: Pagel et al. [14] used a concatenative speech synthesizer and collapsed members of broad phonetic categories to a single instance, e. g. /p/, /t/, and /k/ were mapped to /t/, and all vowels were replaced by /a/ or /ɛ/. Liquids and glides were unchanged. Durations were transformed relative to the average intrinsic phoneme duration. The F0 trajectory remained unmodified. This approach has the potential to create the highest-quality delexicalizations (especially if intrinsic energies and F0 were considered as well). Unfortunately, it relies on a correct phonetic transcription and the availability of a high-quality acoustic inventory of the input speaker, which is not available in most applications.

3. Method

The goal for our proposed algorithm is delexicalization while preserving as much naturalness, speaker identity, and prosodic features as possible, without reliance on formant tracking or speech recognition techniques. The key idea behind our approach is to replace voiced regions with a spectral slice from a *surrogate vowel*, which has been picked from a user-specified location in a training sentence, and to average the spectral contents of unvoiced regions, to render consonants unintelligible. While in our current implementation the selection of the surrogate vowel slice is performed manually, work is underway to automate this step. We now describe the algorithm in more detail.

3.1. Training

The purpose of training is to capture the spectral contents of a user-selected surrogate vowel. The training utterance that contains the desired surrogate vowel is used as input to a pitch-synchronous, harmonic sinusoidal analysis. First, glottal clo-

sure instances (GCIs) are estimated using a center-of-gravity approach [15]. Additional auxiliary markers are created in unvoiced regions at a rate close to a desired fundamental period. The union of GCIs and auxiliary markers form the set of time markers that allow frame-by-frame processing of the entire utterance. Then, each frame $\hat{s}_w[n]$, defined by two consecutive pitch periods, or equivalently the first and last of three time markers, is modeled as

$$\hat{s}_w[n] = w[n] \sum_{l=1}^{L} A_l \sin(\theta_0 n + \phi_l)$$

where $w[n]$ is an asymmetric triangular window, θ_0 is the angular pitch frequency associated with the current fundamental frequency F_0, and $\{A_l, \phi_l\}$ are the sinusoidal magnitude and phase parameters, respectively. For each frame, the sinusoidal parameters are obtained via a Toeplitz set of linear equations [16].

After sinusoidal analysis, the program displays the log-magnitude spectrogram of the utterance, and the user selects the desired vowel spectral slice graphically. Finally, the sinusoidal parameters and the associated F_0 of the frame at the selected time are stored.

3.2. Delexicalization

During delexicalization, an input utterance is first analyzed identically to the procedure during training. For each voiced frame, the magnitude parameters in the frequency band between 200 Hz and 7000 Hz are replaced by the magnitude parameters from the surrogate vowel, after an appropriate interpolation to adjust for the new F_0 of the frame under consideration. Then, the overall root-mean-square energy in that frequency band is adjusted to equal the previously existing energy in that band. The phases remain unmodified. For each unvoiced frame, the magnitude parameters in the frequency band between 400 Hz and 6000 Hz are replaced by the average energy in that band.

As a consequence of this procedure, (1) spectral detail is preserved by modifying only a portion of the frequency band; specifically, important cues related to the voice source are preserved (e. g. open quotient, harmonic-to-noise ratio), (2) short-term (e. g. shimmer) and long-term energy (e. g. word emphasis) fluctuations are preserved, (3) macro (e. g. contrastive or affective prosody) and micro (e. g. jitter) intonation is preserved by means of pitch-synchronous processing without alteration of GCIs, and (4) the impulsive or fricative nature of consonants is preserved by leaving the phase information unmodified and the intrinsic duration of the events themselves. Also, we hypothesize that nasals and approximants retain their perceptual characteristic by virtue of reduced overall energy, compared to the energy of vowels. Figure 1 shows example input and output signal waveforms and spectrograms.

4. Experiment

We carried out several small perceptual listening tests to measure the performance of the proposed algorithm. Specifically, we used an affect recognition test to demonstrate that affective prosody is unaltered, and an intelligibility test to show that intelligibility is practically zero. We used twelve listeners aged 24–60, with self-reported normal hearing. Normalized stimuli were presented over headphones in a quiet room.

	Keyword Recall	Affect Recognition
Delex-classic	0.8%	45.8%
Delex-proposed	0.0%	49.0%
Original	91.6%	65.6%

Table 1: Test results comparing all three stimulus conditions. The first column shows the percentage of correctly recalled keywords. The second column shows the percentage of correctly identified affect.

4.1. Stimulus Conditions

We tested speech utterances in 3 conditions: (1) "delex-classic", a classic delexicalization method using a bandpass filter, as described in Section 2, which will serve as a baseline, (2) "delex-proposed", our proposed delexicalization method, and (3) "original", referring to unprocessed recordings.

4.2. Intelligibility Test

Since the availability of semantic context is an important factor contributing to speech intelligibility, we used conversationally-spoken sentences that were syntactically correct, but semantically anomalous, produced by a single male speaker. Each sentence contained five keywords, e. g. "They *slide far across* the *tiny clock*". They were created by randomizing and exchanging words and grammar structures from the IEEE Harvard Psychoacoustic Sentences. We created 36 unique stimuli using a balanced 12 sentences $\times$ 3 conditions design: each sentence was presented in every condition to four blocks of three listeners. To each individual listener, we presented twelve unique sentences in randomized order. Listeners were asked to repeat a sentence as much as possible, while an administrator counted the number of correctly recalled keywords.

On average, 4.58 keywords per sentence, or 91.6% of all keywords, were recalled correctly in the original condition (see Table 1); this is close to a rate of 92.5% in earlier work [17]. Listeners were, on rare occasion, able to recall single words in the delex-classic condition, which resulted in an average recall rate of 0.04 keywords per sentence, or 0.8%. Not a single keyword was recalled in the delex-proposed condition (see Table 1).

4.3. Affect Recognition Test

For the affect recognition test, we used twelve sentences produced with four affects: angry, fearful, happy, and sad. The speaker was the same as in the previous test. Sentences included affectively-ambiguous exclamations such as "I don't believe it" and "Oh boy". These were elicited using small vignettes [18], for example:

Sad "They were driving past the terrible car accident and saw what had happened. They became both very silent and sad. After a while, she said: *Oh boy*."

Angry "This is just fantastic. Look at the mess you made. You wait and see what dad is going to do. He is gonna let you have it! *Oh boy*!"

Fear "The man was struck with panic. He had just invested the very last of his savings on a risky stock. Then he heard the company might report a massive loss, and he said: That's gonna ruin me. *Oh boy*!"

Happy "I can't believe it, I got an A in math! My dad's gonna give me $25 now because that's what he said. And then I'm gonna buy all these toys! *Oh boy*!"

	Naturalness	Speaker ID
Delex-classic	36%	12%
Delex-proposed	64%	88%

Table 2: Test results comparing the delexicalized stimulus conditions. The first column shows the average preference in percent as to which method sounded more natural. The second column shows the average preference as to which method preserved more speaker identity.

We created 144 unique stimuli using a balanced 12 sentences $\times$ 4 affects $\times$ 3 conditions design: each sentence was presented in every affect and in every condition to twelve listeners. To each individual listener, we presented twelve unique sentences in randomized order. Listeners were played a sentence, and asked to decide which of the four available affects best matched the sentence affect.

On average, 65.6% of affects were recognized correctly in the original condition. Both delex-classic and delex-proposed conditions had lower rates of 45.8% and 49.0%, respectively (see Table 1). The difference between the two delexicalized conditions was statistically insignificant, using a paired t-test.

4.4. Naturalness and Speaker Identity

In a preference test, twelve listeners were asked to state which of the two delexicalized conditions sounded more natural. Of these listeners, 64% preferred the delex-proposed condition, citing that the delex-classic condition sounded "too muffled". The other listeners preferred the delex-classic condition over the "processed" quality of the delex-proposed condition. In a second preference test, 88% of those same twelve listeners judged the delex-proposed condition to preserve more speaker identity than the delex-classic condition (see Table 2).

5. Conclusion

We have presented an algorithm that is a first step towards the goal of delexicalization while preserving, as much as possible, naturalness, speaker identity, and important acoustic-prosodic features. Our algorithm renders an utterance completely unintelligible, while preserving drastically more speaker identity and sounding more natural than the baseline algorithm based on bandpass filtering. Only a modest improvement in affect recognition was achieved compared to the baseline performance.

In the future, we plan on automating the training, as well as studying on how spectral balance can be preserved during delexicalization. Naturalness may be further improved by using a speaker-adapted, Hidden Markov Model that can reliably recognize broad phonetic classes; in this case, each class could be replaced by its associated representative surrogate phoneme.

6. References

[1] K. Atkinson. Language identification from non-segmental cues. *Working Papers in Phonetics*, 10:85–89, 1968.

[2] J. J. Ohala and J. B. Gilbert. Listeners' ability to identify languages by their prosody. In P. Léon and M. Rossi, editors, *Problêmes de Prosodie*, volume 18, pages 123–131. Studia Phonetica, 1979.

[3] Gerit P. Sonntag and Thomas Portele. Looking for the presence of linguistic concepts in the prosody of spoken utterances looking for the presence of linguistic concepts in the prosody of spoken utterances. In *Concept to Speech Generation Systems, Proceedings of a Workshop in conjunction with 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, 1997.

[4] J. Maidment. Voice fundamental frequency characteristics as language differentiators. *Speech and Hearing Work in Progress*, 2:74–93, 1976.

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

[6] B. Blesser. Speech perception under conditions of spectral transformation: I. phonetic characteristics. *Journal of Speech and Hearing Research*, 15:5–41, 1972.

[7] J. Kreimann. Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, 10:163–175, 1982.

[8] I. Lehiste. Perception of sentence and paragraph boundaries. In B. Lindblom and S. Öhman, editors, *Frontiers of speech communication research*, pages 191–201. Academic Press, New York, 1979.

[9] D. Schaffer. The role of intonation as a cue to topic management in conversation. *Journal of Phonetics*, 12:327–344, 1984.

[10] G. Sonntag and T. Portele. PURR - a method for prosody evaluation and investigation. *Computer, Speech and Language*, 12(4):437–451, 1998.

[11] P. Nicolas and P. Roméas. Evaluation of prosody in the french version of a multilingual text-to-speech synthesis: neutralising segmental information in preliminary test. In *Proceedings of Eurospeech*, pages 211–214, Berlin, 1993.

[12] J. R. de Pijper and A. A. Sandermann. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, 96(4):2037–2047, 1994.

[13] M. Vainio, A. Suni, T. Raitio, J. Nurminen, J. Järvikivi, and P. Alku. New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis. In *Proceedings of Interspeech*, Brighton, 2009.

[14] V. Pagel, N. Carbonell, and Y. Laprie. A new method for speech delexicalization, and its application to the perception of french prosody. In *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, 1996.

[15] Y. Stylianou. Removing linear phase mismatches in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, 9(3):232–239, 2001.

[16] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, 9(1):21–29, 2001.

[17] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *Journal of the Acoustical Society of America*, 124(4), October 2008.

[18] E. Klabbers, T. Mishra, and J. van Santen. Analysis of affective speech recordings using the superpositional intonation model. In *Proceedings of the 6th ISCA workshop on speech synthesis (SSW6)*, pages 339–344, 2007.