Analysis and Synthesis of Glottalization Phenomena in German-Accented English

Ivan Kraljevski[1], Maria Paola Bissiri[2], Guntram Strecha[2], and Rüdiger Hoffmann[2]

[1] VoiceINTERConnect GmbH, Dresden, Germany
[2] TU Dresden, Chair for System Theory and Speech Technology, Dresden, Germany
ivan.kraljevski@voiceinterconnect.de
{maria_paola.bissiri,guntram.strecha,
ruediger.hoffmann}@tu-dresden.de

Abstract. The present paper investigates the analysis and synthesis of glottalization phenomena in German-accented English. Word-initial glottalization was manually annotated in a subset of a German-accented English speech corpus. For each glottalized segment, time-normalized F0 and log-energy contours were produced and principal component analysis was performed on the contour sets in order to reduce their dimensionality. Centroid contours of the PC clusters were used for contour reconstruction in the resynthesis experiments. The prototype intonation and intensity contours were superimposed over non-glottalized word-initial vowels in order to resynthesize creaky voice. This procedure allows the automatic creation of speech stimuli which could be used in perceptual experiments for basic research on glottalizations.

Keywords: glottalization, speech perception, speech synthesis.

1 Introduction

In the present paper glottalization is employed as a cover term, defining two major speech phenomena: glottal stops and creaky voice. Glottal stops are produced by closing and abruptly opening the vocal folds. Creaky voice is a more frequent and perceptually equivalent phenomenon, consisting in irregular and low frequency vocal fold vibrations.

In some languages, such as Arabic, glottalizations can be phonemic, i.e. they can differentiate word meaning. This is not the case in German and in English, however in both languages glottalizations are relevant for speech communication. In German, glottalizations are very frequent at word-initial and morpheme-initial vowels [1,2], and are therefore relevant indicators of word and morpheme boundaries. In English glottalization of word-initial vowels is less frequent and more likely to occur at phrase boundaries and pitch accented syllables [3]. German learners of English could transfer their word-linking habit of frequent word-initial glottalization to their English productions, which might therefore sound jerking and overemphasizing to English native speakers [4].

The automatic analysis of glottalizations is seldom carried out [5] because large annotated speech databases and suitable algorithms are rarely available. Acoustic modeling of glottalization can improve Automatic Speech Recognition (ASR) performance [6] since glottalizations can be good cues to word boundaries. Regarding speech synthesis, glottalization modeling is considered useful in order to improve naturalness [5].

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 97–104, 2014.

Furthermore, given the different occurrences and linguistic functions of glottalizations in different languages, their appropriate realization in synthesized speech is desirable.

Inserting a sudden drop in F0 in the target vowel is sufficient to elicit the perception of glottalization [7], however, in order to synthesize glottalizations, it is preferable to manipulate also spectral tilt and the duration of the glottal pulses. HMM-based speech synthesis has been employed also for creaky voice. Csapó and Németh [8] used a synthesis model with three heuristics: pitch halving, pitch-synchronous residual modulation with periods multiplied by random scaling factors and spectral distortion. Raitio et al. [9] presented a fully automatic HMM-based system for synthesis of creaky voice.

In the present paper, pitch and intensity analysis was performed on word-initial glottalizations in a German-accented English corpus, and Principal Component Analysis (PCA) was employed to reduce the dimensionality of the analyzed intonation contours. The component vectors were classified into clusters as glottal stops and creaky voice. The cluster centroids were estimated and used for F0 and log-energy contours reconstruction. The intonational and intensity contours of glottalizations were superimposed on natural unglottalized speech and the acoustic characteristics of the resulting resynthesized speech were evaluated by means of informal listening tests and comparison of voice quality measures.

2 Acoustic Characteristics of Creaky Voice

Glottal stops are characterized by the closure of the glottal folds, visible as a silent phase in the spectrogram, followed by its abrupt opening, after which some irregular vocal fold vibrations at the onset of the following sound can appear.

Creaky voice is a mode of vibration of the vocal folds, in which they are more addicted together. This mode of vibration can affect some more adjacent segments or just part of them, e.g. a single vowel can be realized as partly creaky and partly modal voiced. Creaky voice does not significantly affect the formants of a sound, its more typical characteristics are low F0, reduced intensity and also increased period to period irregularities.

Automatic F0 and intensity measures are not always reliable indicators of creaky voice. F0 detection algorithms are well known to fail in creaky stretches of speech, and intensity can vary for other reasons besides voice quality, e.g. recording conditions or speaker's loudness level. The specific spectral structure of creaky voice can be more useful to detect it than F0 and intensity [6].

For instance, spectral tilt, i.e. "the degree to which intensity drops off as frequency increases" [10], is reported to be more steeply positive for creaky than for modal phonation. Accordingly, in creaky voice the amplitude of the second harmonic (H2) has been found to be higher than the amplitude of the first harmonic (H1) [11].

3 Speech Database

The speech database employed in the present investigation consists in BBC news bulletins read by 4 male and 3 female German native speakers, studio recorded with 44.1 KHz resolution and then downsampled to 16 kHz and 16 bit. It has a total duration of

3 hours and 13 minutes and is composed of 418 recorded sequences. In 102 of them, about 38 minutes of speech, glottalization of word-initial vowels was manually labeled by an expert phonetician (the second author).

The following categories were labeled: absence of glottalization (0), glottal stop (G), creaky voice (CR), breathy voice (H) and sudden drop in F0 (L). Since a glottal closure can be followed by a longer stretch of creaky vowel, the criterion for labeling as glottal stop was that the closure should have a duration of at least 2/3 of the whole glottalization. If the closure was shorter, the segment was categorized as creaky voice.

4 Pitch and Intensity Analysis

In the present investigation, the approach by [12] was employed to analyze pitch in the speech corpus. It is a hybrid pitch marking method that combines outputs of two different speech signal based pitch marking algorithms (PMA) where the pitch marks are combined and represented by Finite State Machine (FSM).

The most accurate pitch marks, those with the highest confidence score, are chosen in the selection stage. The pitch marking was performed for each utterance in the BBC corpus and the results were stored for further analysis.

Mel-cepstral analysis was performed with a frame rate of 5 ms and frame duration of 45 ms, with Blackman window applied, and the frequency band up to 16 KHz with 40 Mel DFT filters. The logarithm of the energy were calculated at each channel output. The zero coefficients presented the log of energy of the analyzed frame and were stored as intensity contours for each utterance for further statistical analysis.

In order to eliminate the speaker dependent variations in F0, energy and segment duration, time normalization was performed on F0 and log-energy contours of the glottalized segments in the BBC corpus. For each glottalization label, the linear time normalization was performed by producing a relative duration expressed as the percentage of the label length (100 samples). Cubic spline interpolation was used in the normalization by smoothing first and second derivatives throughout the curve. Subsequently, the observed maximum was subtracted from the linear time normalized contours. For the speech data employed, the predicted F0 value domain was set between 0-400 Hz. The F0 analysis algorithm also produced negative values indicating unvoiced frames, which were treated as absence of F0 and equaled to 0.

In this way, the mean F0 and log-energy values and their variation over the duration of each glottalized segment can be observed in Tab. 1. This analysis procedure allows the easy manipulation and generation of F0 and intensity contours for the resynthesis of glottalization phenomena.

5 Principal Component Analysis

Principal Component Analysis (PCA) as an orthonormal transformation is widely used in multivariate data analysis for dimension reduction and decorrelation. PCA is carried out to capture most of the variation between the contours using a smaller number of new variables (principal components), where each of these components represents a linear combination of the contour parameters.

Table 1. Mean and standard deviation of F0, intensity and segment duration for word-initial creaky voice (CR), glottal stops (G) and non-glottalized vowels (0)

		F0 in Hz	Intensity in dB	Duration in ms.
label	counts	mean (sd)	mean (sd)	mean (sd)
CR	572	124.62 (63.48)	7.23 (0.65)	80 (31)
G	307	26.76 (25.56)	6.15 (0.72)	90 (31)
0	174	157.30 (52.48)	7.30 (0.60)	58 (19)

The number of components that should be retained is chosen according to the criterion of minimal total variance. After performing PCA on the contour set, it was observed that in both cases retaining the first 5 components is enough to cover more of 90% of the variation (for F0, 90.35% and for the log-energy 98.42%).

Figure 1 presents the relationship between the first two components. Glottalization phenomena – creaky voice, glottal stop and absence of glottalization – are clearly distinguished even by employing only the first two components.

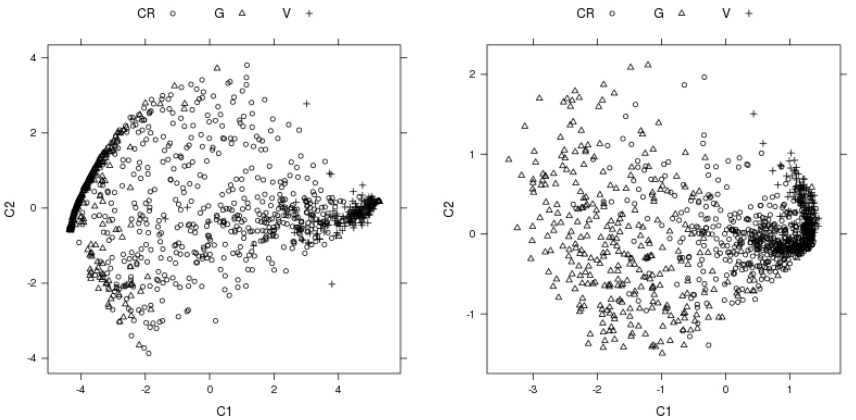
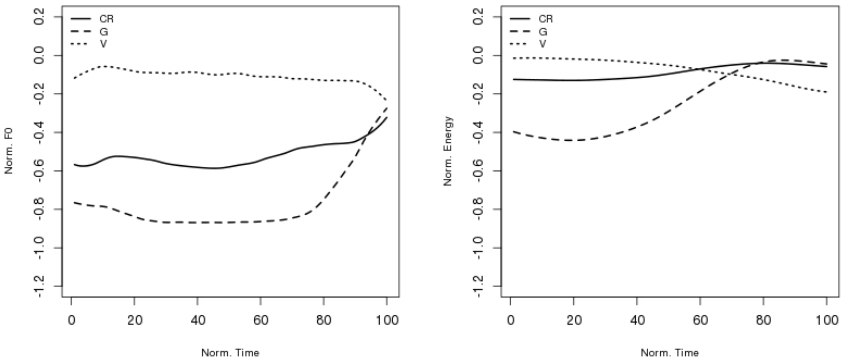


Fig. 1. PC1 vs PC2 for F0 (a) and log-Energy (b) for creaky voice (CR), glottal stops (G) and non-glottalized vowels (V)

6 Synthesis of Glottalization Phenomena

Besides PCA analysis, non-parametrical statistical tests were performed over the F0 and log-energy data sets confirming the existence of distinctive contour features between the observed glottalization phenomena. The analysis of non-glottalized word-initial vowels was included to provide comparison of the acoustic features in the annotated segments. On Fig. 2, the F0 and intensity typical contours are presented.



(a) (b)

Fig. 2. F0 (a) and intensity (b) typical contours for creaky voice (CR), glottal stops (G) and non-glottalized word-initial vowels (V) used for the synthesis experiments

They are reconstructed using the centroid vectors obtained by means of the supervised clustering procedure of the PCA components. The representative vectors of the first five principal components were multiplied with the transposed transformation matrix. It can be seen that the prototype contours express the distinctive features of the glottalization phenomena.

For example, glottal closures larger than 2/3 of the segment duration are a characteristic of the G labels (glottal stops), thus producing a large relative intensity drop in the G prototype contour. In the case of CR, the pitch halving effect with small F0 variation is noticeable, while the intensity reduction is relatively small and constant. For the non-glottalized vowel, F0 and intensity are constant over the whole segment except in the vicinity of the 10% duration from the boundaries, as a result of the consistency in the manual labeling procedure.

Our main motivation to resynthesize glottalization phenomena was to automatically generate stimuli for perception experiments on glottalizations. By means of resynthesis it should be possible to create stimuli that differ only because of the presence or absence of glottalization and are identical in any other aspect. Resynthesis experiments were conducted on short speech segments from the BBC corpus. The F0 and intensity contours were superimposed on a part of an utterance selected from the BBC corpus.

The pitch analysis was performed using the hybrid method described in Sec. 4 with 10 ms shift and 45 ms frame width, and the maximal log-energy value was estimated on the right boundary of the word-initial vowel. Furthermore, in order to increase naturalness, jitter and shimmer were increased by randomly varying timing and amplitude in the synthetic contours according to the following equation:

$$y(n) = x(n) + (-1)^n \cdot a \cdot b(n) \cdot Rand(n), n = 1, ..N. \quad (1)$$

Where *x(n)* is the value for the frame *n* of the prototype contour, *b(n)* is the corresponding value of the standard deviation for a sample and *Rand(n)* is a random number generated from a normal distribution. The intensity was modulated by random values

up to a=0.1 (1 for F0) of the normalized value, since larger ones introduce unwanted distortions. The F0 and intensity contours were superimposed and the resulting synthetic speech was created by generating an excitation pulse train (lower frequencies) and noise (higher frequencies), which were passed through Mel Log Spectrum Approximation (MLSA) filter. In Fig. 3, the spectrograms and the averaged spectra of an example of synthetic speech with the three cases of word-initial vowel glottalizations are presented: absence of glottalization (0), creaky voice (CR), and glottal stop (G).

The duration of the label is chosen to corresponds to the estimated mean duration (80–90 ms) of the CR and G manual labels in the corpus.

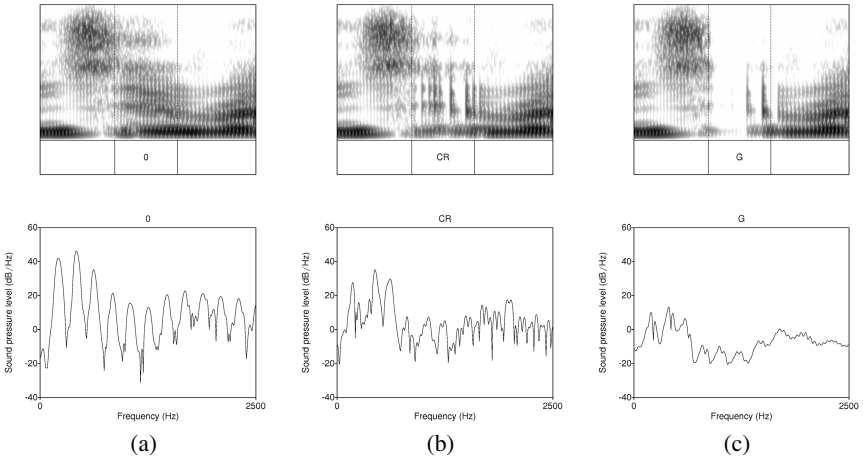


Fig. 3. Spectrograms (top) and averaged spectra (bottom) of a speech segment synthesized in three versions: with word-initial (a) non-glottalized vowel (0), (b) creaky voice (CR), and (c) glottal stop (G).

The synthesized creaky voice exhibits acoustic characteristics typical for natural creaky voice: lower energy as well as steeper spectral slope (H2-H1) than modal voice; irregular periods are also visible in the spectrogram. The synthetic glottal stop shows very low energy in its first part and a couple of irregular glottal pulses in its second part, as it occurs in natural glottal stops.

In order to support the observations from the inspection of the spectrogram, voice quality analysis was conducted using Praat. Jitter, shimmer, mean harmonics-to-noise ratio (MHNR), H1 and H2 were measured in the original non-glottalized signal, and in the corresponding synthetic signals with non-glottalized vowel, creaky voice and glottal stop (see Tab. 2). The original non-glottalized segment has increased jitter (3.2%) compared to the synthetic non-glottalized segment which indicates that the vowel in this case is produced with lower quality. The reason is the influence of the the preceding consonant, which was reduced after the resynthesis. Moreover, for accurate measurements of vowel quality parameters, it is recommended to analyze longer vowel segments.

The non-glottalized synthetic segment, thus without F0 and intensity modification, introduces lower jitter, increased shimmer and equal MHNR ratio compared to the original one. The synthetic CR has much lower values for HNR and increased shimmer,

Table 2. Voice quality analysis for the original non-glottalized vowel and for the three corresponding synthetic signals: non-glottalized vowel, creaky voice and glottal stop

	Jitter (%)	Shimmer (%)	MHNR (dB)	H2-H1 (dB)
Original (non-synthetic)	3.2	6.9	11.8	1.5
Non-glottalized vowel (0)	2.3	10.4	11.8	4
Creaky voice (CR)	2.9	21.3	6.1	7.4
Glottal stop (G)	1.1	18.2	5.8	2

which applies to the G segment as well. Informal listening tests also indicated that the synthetic glottal stops and creaky voiced vowels were easy to identify as such and sounded similar to natural ones.

The approach proposed in the present paper can effectively manipulate stretches of modal voice transforming them into glottalizations. It is just necessary to define the start and end points of the segments that need to be glottalized. These should be long enough – e.g. around 80 ms, the mean duration of G and CR labels in the corpus (see pag. 102) – otherwise too much downsampling of the F0 and intensity prototype contours would not deliver a good synthesis. This is especially valid for the synthesis of glottal stops since the glottal closure cannot be realized properly if the segment is too short. For segments of sufficient length the manipulation delivers good quality glottalizations without artifacts.

By informally comparing manipulated utterances with and without glottalization, the impression is often that the glottalization inserts a stronger phrase boundary or emphasizes the target word, as it occurs with natural glottalizations in English. The proposed method can thus create speech stimuli suitable for investigating the influence of glottalizations on speech perception.

7 Conclusions

In the present paper the analysis and synthesis of glottalization phenomena in German-accented English are presented. Glottalizations of word-initial vowels were manually annotated by a human expert in a small part of a German-accented English speech corpus. Pitch and intensity analysis were performed on the annotated subset of the corpus, and for each word-initial segment, labeled as glottal stop, creaky voice or non-glottalized vowel, time normalized F0 and intensity contours were produced. The contours describe the relative variation compared to the maximal observed values in the segment. Multivariate statistical analysis was performed on the contour data sets in order to find the principal components and to reduce the contour dimensionality (from 100 to 5). Clustering analysis of the PCs gave the centroid contours which were used in the resynthesis experiments. The prototype contours were modulated with random values in order to simulate the effects of jitter and shimmer. Such F0 and log-energy contours were superimposed on natural non-glottalized word-initial vowels chosen from the corpus. Qualitative analysis (spectrum observations and informal listening tests) as well as quantitative analysis (voice quality measurements) indicated that the synthetic utter-

ances indeed possess the desired glottalization phenomena characteristics. This procedure could be employed for the automatic generation of speech stimuli for perceptual experiments on glottalizations.

References

1. Kohler, K.J.: Glottal stops and glottalization in German. Data and theory of connected speech processes. Phonetica 51, 38–51 (1994)
2. Kiessling, A., Kompe, R., Niemann, H., Nöth, E., Batliner, A.: Voice source state as a source of information in speech recognition: detection of laryngealizations. In: Speech Recognition and Coding. New Advances and Trends, pp. 329–332. Springer, Berlin (1995)
3. Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M.: Glottalization of word-initial vowels as a function of prosodic structure. Journal of Phonetics 24, 423–444 (1996)
4. Bissiri, M.P.: Glottalizations in German-accented English in relationship to phrase boundaries. In: Mehnert, D., Kordon, U., Wolff, M. (eds.) Systemtheorie Signalverarbeitung Sprachtechnologie, pp. 234–240. TUD Press, Dresden (2013)
5. Drugman, T., Kane, J., Gobl, C.: Modeling the creaky excitation for parametric speech synthesis. In: Proc. of Interspeech, Portland, Oregon, pp. 1424–1427 (2012)
6. Yoon, T.-J., Zhuang, X., Cole, J., Hasegawa-Johnson, M.: Voice quality dependent speech recognition. In: Proc. of Int. Symp. on Linguistic Patterns in Spontaneous Speech, Taipei, Taiwan (2006)
7. Pierrehumbert, J.B., Frisch, S.: Synthesizing allophonic glottalization. In: Progress in Speech Synthesis, pp. 9–26. Springer, New York (1997)
8. Csapó, T.G., Németh, G.: A novel irregular voice model for HMM-based speech synthesis. In: Proc. ISCA SSW8, pp. 229–234 (2013)
9. Raitio, T., Kane, J., Drugman, T., Gobl, C.: HMM-based synthesis of creaky voice. In: Proc. Interspeech, pp. 2316–2320 (2013)
10. Gordon, M., Ladefoged, P.: Phonation types: A cross-linguistic overview. Journal of Phonetics 29(4), 383–406 (2001)
11. Ni Chasaide, A., Gobl, C.: Voice source variation. In: Hardcastle, W.J., Laver, J. (eds.) The Handbook of Phonetic Sciences, pp. 427–461. Blackwell, Oxford (1997)
12. Hussein, H., Wolff, M., Jokisch, O., Duckhorn, F., Strecha, G., Hoffmann, R.: A hybrid speech signal based algorithm for pitch marking using finite state machines. In: INTERSPEECH, pp. 135–138 (2008)