HMM-BASED CLASSIFICATION OF GLOTTALIZATION PHENOMENA IN GERMAN-ACCENTED ENGLISH

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

[1]TU Dresden, Chair for System Theory and Speech Technology, Dresden, Germany
[2]VoiceINTERConnect GmbH, Dresden, Germany

{maria_paola.bissiri, ruediger.hoffmann}@tu-dresden.de
ivan.kraljevski@voiceinterconnect.de

Abstract. The present paper investigates the automatic detection of word-initial glottalization phenomena (glottal stops and creaky voice) in German-accented English by means of HMMs. Glottalization of word-initial vowels can be very frequent in German-accented English, as well as in German. Detection and classification of glottalization phenomena is useful in order to obtain a pre-segmentation of speech for basic phonetic research. Moreover, the inclusion of acoustic models for glottalization, which are good indicators of word and morpheme boundaries, can improve the overall performance of an Automatic Speech Recognition (ASR) system. Since creaky voiced vowels have different spectro-temporal acoustic characteristics than modal voiced vowels, it is possible to train HMM models for creaky voice. A corpus of BBC news bulletins read by German speakers was recorded and a subset of the corpus with manually labeled glottalization at word-initial vowels was used to initialize HMM models. These models were used to label the training and the test set in the corpus. To analyze detection accuracy, the automatically derived labels were compared with the manual labels. Results show that the proposed framework can be used for the identification of glottalization phenomena, providing a reliable automatic pre-segmentation for the purpose of phonetic basic research.

1. Introduction

Glottalizations are produced by opening and then abruptly closing the vocal folds (glottal stops) or by means of irregular and low frequency glottal fold vibrations (creaky voice). In German and in English, glottalizations are not phonemes, i.e. they cannot differentiate word meaning, however, they can be quite relevant in speech communication. Glottalizations can carry out non-verbal information about speaker's dialect, attitude, mood and emotional state [1, 2]. Although they are produced involuntarily, glottalizations can also carry out relevant linguistic information by marking morpheme, word, phrase, or utterance boundaries, depending on the language. For instance, in German, glottalizations of word-initial vowels are frequent at word and morpheme boundaries [3, 4]. In English glottalizations often occur at intonation phrase boundaries as well as at pitch accented syllables [5], in Swedish [6] they are phrase boundary markers, in Finnish and in English markers for turn taking [7, 8].

The automatic identification of glottalization phenomena is difficult because large annotated speech databases and suitable automatic detection algorithms are seldom available. As a result, several studies report qualitative analysis of glottalizations and less investigations are concerned with their automatic detection [9]. However, the inclusion of acoustic models for glottal stops and creaky voice in Automatic Speech Recognition (ASR) can have several advantages. Robust automatic detection of glottalized segments would provide a useful pre-segmentation for basic phonetic research. Since glottalizations can be good indicators of word boundaries, for instance in German and in Czech [10], the acoustic modeling of glottalizations can improve ASR overall performance [11].

The parametrization of creaky voice in the speech signal could be implemented in the automatic recognition of tones since creaky voice can contribute to tone perception [12]. In automatic pitch detection, glottalized segments introduce errors because of their irregular characteristics (see Section 2), therefore their identification and exclusion would improve pitch analysis. Moreover, the detection of creakiness could be also exploited for improving speaker recognition systems. Regarding speech synthesis, glottalization modeling can contribute to the naturalness of synthesized speech [9].

Various approaches have been proposed to identify creaky voice automatically, for instance, based on autocorrelation [13], HMM models [14, 15], pitch marking [16], narrow band Fourier spectrum analysis [17], and on Linear Prediction (LP) residual based features [18]. In a previous investigation on German-accented English [19], an existing ASR system with German HMM models including the glottal stop was used for the automatic detection of glottal stops before word-initial vowels in order to create a pre-segmentation for the purpose of basic phonetic research. The automatic labels were compared with the manual ones and it was observed that glottal stops were reliably identified. The additional detection of creaky voice is desirable since it has been found as the most frequent glottalization phenomenon at word-initial vowels in German [3] and in German-accented English [10]. In order to detect creaky voice, additional acoustic modeling is required.

2. Acoustic characteristics of glottal stops and creaky voice

Glottal stops are plosives characterized by a complete closure of the vocal folds, indicated by a phase of silence in the spectrogram (s. Figure 1a). The subsequent glottal fold opening can be followed by some irregular vocal fold vibrations, also visible in the spectrogram.

Creaky voice is a mode of vibration of the vocal folds in which they are closer together. This mode of vibration can affect whole voiced segments or part of them, while it does not significantly change their formant characteristics. Typical for creaky voice are reduced intensity, low F0 (e.g. below 50 Hz) and increased period to period irregularities with jitter values over 10% [17]. However, since F0 detection algorithms often fail in glottalized regions of speech, F0 measures are not always reliable indicators of creaky voice. Intensity could be influenced by external factors other than glottalization, for instance by recording conditions, such as microphone location and environmental noise, as well as by internal factors, such as the speaker's loudness level.

Creaky voice is characterized by a specific spectral structure, which could be more reliable to detect it than F0 or intensity alone [11]. For instance, an acoustic parameter for identifying creaky voice is spectral tilt, defined as "the degree to which intensity drops off as frequency increases" [20]. Spectral tilt can be measured by comparing the amplitude of F0 to that of higher frequency harmonics and it is more steeply positive for creaky voice [20]. Accordingly, in creaky phonation the amplitude of the second harmonic (H2) has been reported to be higher than the amplitude of the first harmonic (H1) [21].

Due to the distinctive spectral characteristics of creaky voice, the processing and classification algorithms commonly employed for speech and speaker recognition can be used for its detection. With the choice of an appropriate feature extraction method, creaky voice can be acoustically modeled using HMMs. Perceptual linear prediction (PLP) coefficients were used in ASR by [8], and it was observed that such features can effectively encode voice quality variations determined by their harmonic and temporal measures.

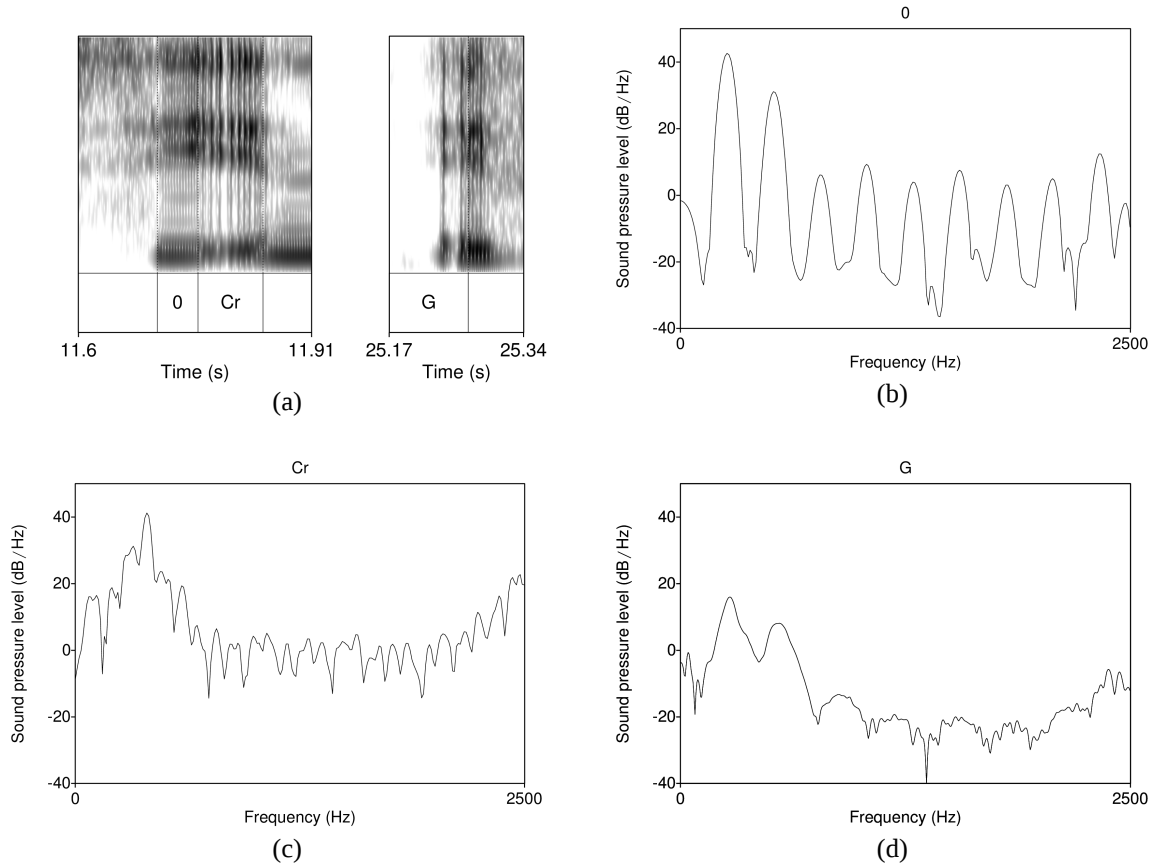

Figure 1. a) spectrogram showing (from left to right) a nonglottalized vowel (0), creaky voice (Cr), and a glottal stop (G), b) averaged spectrum of the nonglottalized vowel (0), c) averaged spectrum of the creaky segment (Cr), d) averaged spectrum of the glottal stop (G).

Lugger et al. [14] found ergodic HMM models with glottal gradients as voice quality features more suitable for classifying breathy, modal, rough and creaky voice qualities. Kane et al. [22] proposed a combination of commonly used spectral features with a specific set of voice quality features, such as the difference between the first two harmonics (H1-H2), for voice quality classification. In general, any feature extraction method capable of capturing the spectral envelope (the harmonic amplitudes) as well as its temporal variation can be successfully applied.

3. Experimental setup

3.1 Speech data

The speech data employed in the present study consists of BBC news bulletins read by 4 male and 3 female German native speakers. The data was studio recorded with 44.1 KHz PCM and then downsampled to 16 KHz and 16 bit resolution. The resulting speech database of 3 hours and 13 minutes duration consists of 418 recorded sequences. The 102 sequences (about 38 minutes) in which glottalization of word-initial vowels was manually labeled by an expert phonetician (the first author) were divided into a training set (86 sequences, TRAIN1), used for creating the initial acoustic models, and a test set (16 sequences, TEST1), used for their validation. The initial models were then employed for the automatic labeling of the whole database. The 316 sequences of the database which were not manually labeled were used for the training of new acoustic models (TRAIN2). The detection accuracy of word-initial glottalization by means of the new models was evaluated on the 102 manually labeled sequences (TEST2).

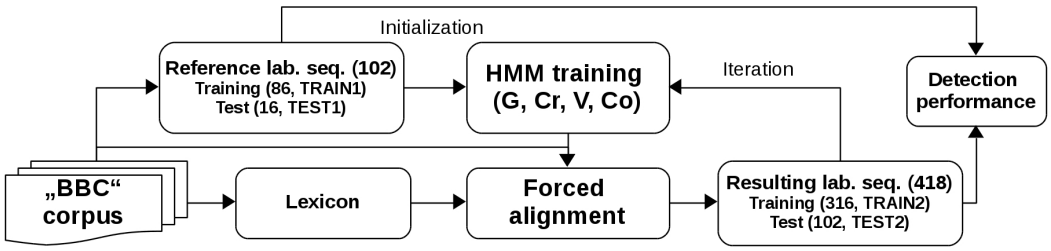

Figure 2. HMM modeling and automatic labeling procedure

3.2 Lexicon

The pronunciation lexicon was created by modifying the one used in [19, 23]. This included only German phonemes obtained through English to German phoneme mapping. The lexicon was enhanced by means of rules reproducing several German-accented English pronunciation variations. The rules were derived from linguistic knowledge of German-accented English (knowledge-based approach), as well as from the analysis of the ASR confusion matrix (data-driven approach) [23]. Because of the limited amount of speech data available, it was not possible to reliably train separate models for all phonemes of the German language. Therefore, differently than in the lexicon used in [19, 23], all vowels and consonants were mapped to the single categories V (for vowel) and Co (for consonant), in order to preserve the context of glottalization phenomena. Word-initial vowels were coded according to the following optional variants: a) V, meaning a nonglottalized vowel, b) G V, a glottal stop preceding the vowel, c) Cr V, an initially creaky and then modal voiced vowel, d) Cr, a completely creaky vowel.

3.3 Glottalization labeling procedure

Glottalization phenomena of word-initial vowels were manually annotated according to the following categories: absence of glottalization, glottal stop, creaky voice, breathy voice, and low F0. Since a glottal stop closure could be followed by a long creaky segment, the criterion for labeling glottal stops was that the glottal closure should be at least 2/3 of the whole glottalization (see [19, 10] for further details on the labeling procedure). The manual segments for word-initial nonglottalized vowels, glottal stops and creaky voice were superimposed over the automatically obtained phone segments as in [23], producing the reference labels. Breathy voice and low F0 labels, which were less frequent in the corpus [10], were not considered.

The reference annotations were used to train the initial HMM models, which were then employed for the automatic labeling of the non-annotated part of the speech database. The labeling was carried out by means of Viterbi forced alignment, using the orthographic transcriptions and the coding in the lexicon (Figure 2). The whole automatically annotated database was used for training new and more accurate HMM models.

3.4 HMM recognizer

The feature extraction was performed by means of a Blackman window applied over 25 ms wide frames with a frame period of 10 ms. The band from 300 to 8000 Hz was covered with 31 Mel DFT filters and the log of the energy was computed at the output of each channel.

The obtained feature vectors and their delta values were normalized to a mean of zero and a standard deviation of one. Principal Component Analysis (PCA), as an orthonormal transformation which provides a linear mapping for dimension reduction and decorrelation, was used to bring the number of components to 24.

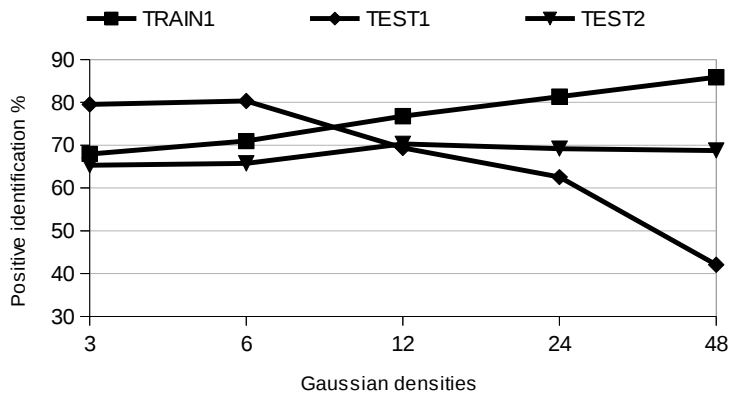

Figure 3. Positive identification of the manual with the automatic labels for word-initial glottal phenomena (averaged values), for the TRAIN1 and the TEST1 sets after the first training process and for the TEST2 set after the second training process across different Gaussian densities.

The detection system employed was implemented by means of the UASR (Unified Approach for Speech Synthesis and Recognition) framework [24]. The system uses arc-emission HMMs with a single Gaussian density per arc. The choice of the feature extraction method was based on previous recognition experiments with the same framework. In these experiments, the Mel-Log-Spectrum performed similarly as MFCC in speaker recognition [25] or better than MFCC in phoneme recognition [24].

The structure is built by an iterative training process by means of state splitting from the initial HMM models. The two glottalization phenomena (Cr) and (G) along with vowel (V) and consonant (Co) were modeled (see Section 3.2). The topology of the initial models is the traditional left-to-right 3-state HMM with one Gaussian density per state.

The number of final states (Gaussian densities) was achieved by state splitting, effectively doubling the number of Gaussians for each iteration. Models produced after each state split were included in the detection experiments and the best performing model on the TEST1 set was chosen for the automatic annotation of the unlabeled part of the speech database (TRAIN2).

4. Results and Discussion

In order to estimate the quality of the obtained HMM models, the automatic were compared with the manual labels. For each manual label of word-initial glottal stop or creaky voice, it was checked for the presence of the corresponding automatic glottalization label in the same word. If there was a match, it was counted as a positive identification. In the case of a nonglottalized word-initial vowel in the manual labeling, if there was no match with an automatic glottalization label in the same word, this was counted also as positive identification.

Figure 3 shows the averaged positive identification of glottal stops, creaky voice and absence of glottalization at word-initial vowels. The presented results are obtained from the TRAIN1 and TEST1 sets (respectively 86 and 16 sequences), and from the TEST2 set (102 sequences) after the second training process, using models with different numbers of Gaussian densities. The HMM models with 12 Gaussian densities provided optimal detection performance on both TRAIN1 and TEST1 set, hence, they were chosen to perform the automatic labeling on the unlabeled part of the speech database. A second training process was initiated using the now automatically labeled TRAIN2 set (316 sequences), and the manually annotated TEST2 set.

Table 1. Positive identification in % of the manual with the automatic labels for word-initial glottal stops (G), creaky voice (Cr) and absence of glottalization (0) on the TEST2 set across different Gaussian densities.

Observed labels	Gaussian densities				
	3	**6**	**12**	**24**	**48**
G	66.9	61.4	67.9	64.0	60.1
Cr	55.9	62.7	67.0	65.9	67.0
0	73.1	73.1	76.1	77.7	79.2

Table 1 shows the positive automatic identification of word-initial glottal stops (G), creaky voice (Cr) and absence of glottalization (0) in the TEST2 set for each training iteration, in comparison with the manual labels. For each word-initial category the best detection performance is obtained for the HMM models with 12 Gaussian densities. The ASR experiments with the 12-Gaussian-models produced the following recognition performance: frame label correctness (FSC) of $83.30 \pm 0.9\%$, label sequence correctness (LSC) of $85.50 \pm 1.4\%$, label sequence accuracy (LSA) of $83.40 \pm 1.2\%$ and lattice density of 0.908 ± 0.018 on the reference annotations.

Table 2 presents the confusion matrix between the reference and the recognized frame labels for creaky voice (Cr), glottal stops (G), consonants (Co) and vowels (V) after the second training process. Confusions are calculated by means of the Levenshtein distance algorithm, which considers deletions, substitutions and insertions. Frames labeled as glottal stop and creaky voice are correctly identified in 70.8% and 72.5% of the cases respectively. The most notable confusions concerning glottalization reference frames are those of creaky voice with the vowel model (13.2%) and of glottal stops with creaky voice (11.1%), followed by glottal stops with vowels (8.6%) and creaky voice with the consonant model (8.5%).

The confusion of creaky voice with vowel frames can be expected since the manual labels for creaky voice obviously refer to glottalized vowels. Moreover, creakiness can affect also vowels which are not in word-initial position, especially those at the end of phrases and utterances [26], therefore some creaky non-initial vowels were included in the training of the HMM vowel model. The confusion of glottal stop with creaky voice, and also of creaky voice with glottal stop (5.8%), can be explained by the labeling criterion used to distinguish the two phenomena, according to which creaky voice could include a glottal closure of less than one third of the whole glottalization (see Section 3.3). This criterion was chosen to ensure comparability of results with previous investigations on Czech-accented English (see [10]). However, the fact that both creaky voice and glottal stops could include a silent part surely makes the automatic identification of their frames more difficult.

The confusion of glottal stop frames with the vowel model could be due to the first irregular vowel periods after the glottal closure, which were included in the manual glottal stop label. The confusion of glottal stop frames with consonants (5.6%) can be motivated by the fact that glottal stops are plosive consonants and thus share some characteristics with other sounds in the HMM consonant model. In general, the confusion with the vowel and consonant models is understandable since the mapping of vowel and consonant sounds to the two categories could not generate precise models. However this mapping proved to be useful, making a good recognition performance possible even with a limited amount of data. The presented results confirm that standard HMM modeling, commonly employed in speech and speaker recognition, can be used for the classification of word-initial glottalization phenomena. The proposed methodology could be also employed for detecting glottalization in other languages besides German-accented English.

Table 2. Frame labels confusion matrix in % (counts in brackets) after recognition evaluation on the TEST2 set using the model with 12 Gaussian densities.

Observed frames	Reference frames				
	silence	**Cr**	**V**	**Co**	**G**
silence	94.2 (76570)	0.0 (0)	0.8 (531)	2.7 (3074)	4.0 (111)
Cr	0.3 (284)	72.5 (3301)	5.2 (3635)	1.9 (2197)	11.1 (306)
V	0.7 (579)	13.2 (602)	79.9 (55705)	19.0 (21465)	8.6 (236)
Co	2.4 (1965)	8.5 (385)	13.3 (9259)	73.8 (83593)	5.6 (154)
G	2.4 (1921)	5.8 (262)	0.9 (601)	2.5 (2868)	70.8 (1953)

5. Conclusions

The present paper investigates the automatic detection of word-initial glottalization phenomena, glottal stops and creaky voice, in German-accented English by means of HMMs. By using a limited amount of data (about 38 minutes of speech) in which word-initial glottalizations had been manually labeled, it was possible to initialize HMM models for glottal stops and creaky voice. These models could be used to label the rest of the German-accented English database.

To analyze detection accuracy, the labels derived by forced alignment were compared to the reference labels, in which word-initial glottalization categories had been labeled by an expert phonetician. HMM models performed well in the detection of creaky voice and glottal stops, which were correctly identified in the majority of cases (respectively 72.5% and 70.8% identification with reference frame labels). Most frame confusions could be due to the vowel- and consonant-character of creaky voice and glottal stops, to the manual labeling criteria or to the presence of one single HMM model for consonants (Co) and one for vowels (V). However, Co and V proved to be useful to model the context of word-initial vowels given the small size of the corpus.

The results of the detection experiments show that the proposed method can be successfully employed for the identification of glottalization phenomena, thus creating a reliable pre-segmentation useful for phonetic basic research. The acoustic models for glottalization phenomena could be further improved by means of re-training or adaptation on larger automatically labeled speech data, thus improving detection accuracy.

6. Acknowledgments

The authors are thankful to Sören Wittenberg for his comments on a previous version of this paper.

7. References

[1] Gobl, C. and Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, 189-212.

[2] Klasmeyer, G. and Sendlmeier, W. F., 2000. Voice and emotional states. In: Kent, R., Ball, D., Martin, J. (eds.), Voice quality measurement. San Diego, CA: Singular/ThomsonLearning, pp. 339–357.

[3] Kohler, K. J., 1994. Glottal stops and glottalization in German. Data and theory of connected speech processes. Phonetica 51, 38-51.

[4] Kießling, A., Kompe, R., Niemann, H., Nöth, E., Batliner, A., 1995. Voice source state as a source of information in speech recognition: detection of laryngealizations, In: Speech recognition and coding. New advances and trends, Berlin: Springer-Verlag, pp. 329-332.

[5] Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M., 1996. Glottalization of word-initial vowels as a function of prosodic structure. Journal of Phonetics 24, 423-444.
[6] Fant, G., and Kruckenberg, A., 1989. Preliminaries to the study of Swedish prose reading and reading style. STL-QPSR 30(2), 001-080.
[7] Ogden, R., 2003. Voice quality as a resource for the management of turn-taking in Finnish talk-in-interaction. In: Proc. of 15th Int. Congr. of Phonetic Sciences, Barcelona.
[8] Local, J. and Kelly, J., 1986. Projections and ‘silences’: notes on phonetic and conversational structure. Human Studies 9, 185-204.
[9] Drugman, T., Kane, J., Gobl, C., 2012. Modeling the creaky excitation for parametric speech synthesis. In: Proc. of Interspeech, Portland, Oregon, pp. 1424-1427.
[10] Bissiri, M.P., 2013. Glottalizations in German-accented English in relationship to phrase boundaries. In: Mehnert, D., Kordon, U., Wolff, M. (eds.), Systemtheorie Signalverarbeitung Sprachtechnologie, Dresden: TUDpress, pp. 234-240.
[11] Yoon, T.-J., Zhuang, X., Cole, J., Hasegawa-Johnson, M., 2006. Voice quality dependent speech recognition. In: Proc. of Int. Symp. on Linguistic Patterns in Spontaneous Speech, Taipei, Taiwan.
[12] Yu, K. M. and Lam, H. W., 2011. The role of creaky voice in Cantonese tonal perception. In: Proc. of 17th Int. Cong. of Phonetic Sciences, Hong Kong, pp. 2240-2243.
[13] Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., Hagita, N., 2008. A method for automatic detection of vocal fry. Audio, speech, and language processing, IEEE Trans. 16(1), 47-56.
[14] Lugger, M., Stimm, F., Yang, B., 2008. Extracting voice quality contours using discrete hidden Markov models. In: Proc. of Speech Prosody 2008, Campinas, Brazil, pp. 29-32.
[15] Cullen, A., Kane, J., Drugman, T., Harte, N., 2013. Creaky voice and the classification of affect. In: Proc. of Workshop on Affective Social Speech Signals, Grenoble, France.
[16] Hagmüller, M. and Kubin, G., 2006. Poincaré pitch marks. Speech Communication 48(12), 1650-1665.
[17] Martin, P., 2012. Automatic detection of voice creak. In: Proc. of Speech Prosody, Shanghai.
[18] Drugman, T., Kane, J. and Gobl, C., 2012. Resonator-based creaky voice detection. In: Proc. of Interspeech, Portland, Oregon.
[19] Bissiri, M.P., Kraljevski, I., Hoffmann, R., 2013. Glottal stop detection in German-accented English using ASR. In: Proc. of AIA-DAGA 2013, Meran, Italy.
[20] Gordon, M. and Ladefoged, P., 2001. Phonation types: a cross-linguistic overview. Journal of Phonetics 29(4): 383-406.
[21] Ni Chasaide, A. and Gobl, C., 1997. Voice source variation. In: Hardcastle, W.J., Laver, J. (eds.), The Handbook of Phonetic Sciences, Oxford: Blackwell, pp. 427-461.
[22] Kane, J., Scherer, S., Aylett, M.P., Morency, L.-P., Gobl, C., 2013. Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech. In: Proc. of ICASSP 2013, 26-31 May 2013, pp. 7982-7986.
[23] Bissiri, M.P., Kraljevski, I., Hoffmann, R., 2013. Improved phoneme segmentation of German-accented Englsih by means of lexicon and acoustic model adaptation. In: Elektronische Sprachsignalverarbeitung 2013, Dresden: TUDpress, pp. 231-238.
[24] Hoffmann, R., Eichner, M., Wolff, M., 2007. Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system. In: Esposito, A. et al. (eds.), Verbal and nonverbal communication behaviours, Berlin, Heidelberg: Springer, pp. 200-218.
[25] Kraljevski, I., Bissiri, M.P., Hoffmann, R.: "Text independent speaker identification with coded speech". In: Elektronische Sprachsignalverarbeitung 2013, Dresden: TUDpress, pp. 239-246.
[26] Redi, L. and Shattuck-Hufnagel, S., 2001. Variation in the realization of glottalization in normal speakers. Journal of Phonetics, 29(4), 407-429.