Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics

Analysis and modeling of speech produced with irregular phonation

Ph.D. thesis booklet

Tamás Mihály Bőhm, M.Sc.

Supervisors:
Géza Németh, Ph.D.
Gábor Olaszy, D.Sc.

Budapest, 2009

1. Introduction

For a long time, it has been a desire to create speaking machines or to communicate with machines via voice. In the past decades, the considerable success of speech technology rendered these aims more realistic. Systems that are capable of reading texts with naturalness close to human speech [1;2] or that can write down spoken sentences [3], with some limitations, are now widely available for several languages.

These successful developments were in part enabled by the fact that researchers focused on 'idealized' human speech. Idealized speech refers to prototypical, young or middle-aged adult, healthy speakers' utterances of grammatical sentences that conform to the generally accepted principles of speech science (e.g. the source-filter theory, or the short-term periodicity of voiced speech). Speech synthesizers and speech recognizers were mainly prepared to process idealized speech (i.e. they are designed with the assumption that the speech signal is ideal).

However, I believe that designing for idealized speech limits further development, as in practice, the speech signal is usually not idealized. Even if we disregard elderly and child speech, as well as pathological speech, the assumptions about idealized speech often become invalid. The novel results of speech science are showing the limitations of these principles.

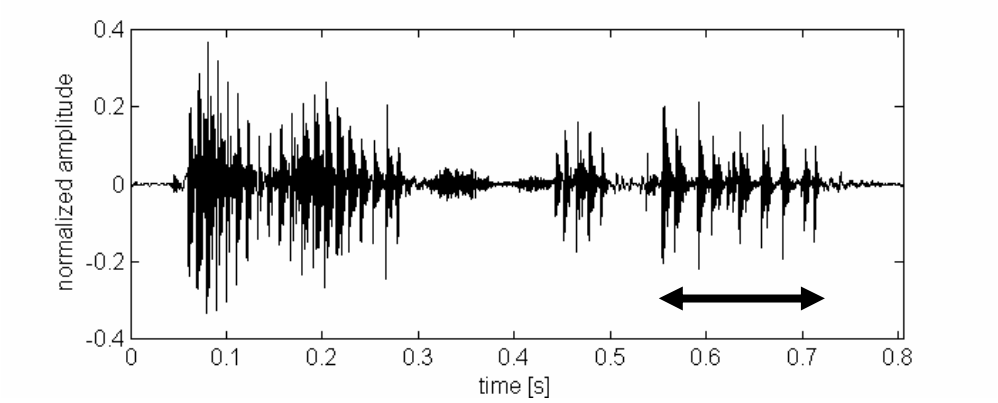

Figure 1: Irregular phonation in a speech waveform
(irregular region denoted by the arrow)

In speech technology, vocal fold vibration was traditionally assumed to be close to periodic, or regular (showing only small cycle-to-cycle changes in the amplitudes and durations of the periods). However, the vibration can deviate from that idealized model: for example, it can become irregular that results in a rough- or harsh-sounding voice quality. Although the phenomenon of *irregular phonation* has been known for a long time, speech technologies applied today usually do not handle it. The reason behind this is probably that, earlier, irregular phonation was considered to be rare and negligible in normal speech. However a growing body of evidence suggests that this phonation type is relatively frequent and that it is likely to have several linguistic and nonlinguistic communicative functions (in certain languages). Such functions include

the acoustic marking of prosodic structure and some emotional states, such as the contribution to the speaker's individual speech characteristics.

Thus appropriate handling of irregular phonation in normal speech has the potential to contribute to several fields of speech technology. For example, the appropriate application of this phonation type in the prosody generation stage of speech synthesis can help to produce more natural sounding or expressive speech. The automatic detection of irregular phonation can improve speech recognition by splitting the input to intonational phrases and by enabling to extract further prosodic information, as well as making it possible to recognize some emotional states. The phonation type differences across speakers can be exploited in speaker identification and voice conversion systems.

Methods for the analysis and synthesis of irregular phonation by machine are a prerequisite for the applications mentioned above. In this dissertation I present such methods: Thesis Group I is about an algorithm that can detect irregular phonation, while in Thesis Group II, I introduce a procedure that is capable of transforming regular voice to irregular voice. In Thesis Group III, I supplement these methods with results on the occurrence and memory storage of speaker differences in phonation types that raise the possibility of further applications.

1.1. Background

In this work, irregular phonation refers to the phenomenon when, due to the irregular vibration of the vocal folds, the speech waveform exhibits abrupt, substantial cycle length and amplitude changes (thus the deviation from periodicity is beyond normal jitter and shimmer[1] [4]; Figure 1) and this deviation is clearly audible for people with normal hearing. It is also considered irregular phonation when the fundamental frequency abruptly drops below the speaker's characteristic voice register and that results in a perceivable change in voice quality.

In the literature, it is common practice to give qualitative definitions of irregular phonation (as the one above) [5-7], because it is not yet clear what combinations of acoustic parameters (changes in jitter, shimmer, F0, amplitude, open quotient and spectral tilt, etc.) define this phonation type.

There are a number of terms used for irregular phonation, both in English (e.g. creaky voice, vocal fry, pulsed phonation, laryngealization, glottalization) and in Hungarian ("recsegő zönge" /raspy voice/ [8]; "érdes, rekedtes hang" /rough, hoarse voice/ [9]; "nyikorgó zönge" /creaky voice/ [10]; "laringalizált, csikorgó beszéd" /laryngealized, screeching speech/ [11]; "nyekergésféle" /scraping/ [12]). Irregular phonation can be a symptom of vocal fold disorders (e.g. asymmetry, nodules, paralysis) [8;13]. However this work was carried out on speech produced with normal, healthy vocal folds. Irregular phonation occurs in normal speech, but, in contrast to pathological speech, it is only intermittent and, as a growing body of results suggest, it may contribute to certain linguistic messages (segmental and suprasegmental) and to non-linguistic information (e.g. emotional state) in speech.

[1] Jitter and shimmer are measures of perturbation in fundamental frequency and amplitude, respectively. Even regular voice is not perfectly periodic, thus there is always some small-scale variation in fundamental frequency and amplitude.

Automatic detection of irregular phonation: For the classification of phonation types (regular vs. irregular), several different approaches can be found in the literature (e.g. [4;7]). It is common in these approaches that, after splicing the input into smaller frames, a set of acoustic cues are extracted from the speech signal (feature extraction), and then the decision is made based on these cue values. Although earlier classifiers use a wide variety of cues and decision algorithms, the performance they achieved (in terms of hit rate and false alarm rate) is not sufficient for many practical applications. Methods used to evaluate each cue (comparison of means and histograms) are either based on assumptions that are likely to be violated (e.g. unimodal distribution) or cannot provide a quantitative performance measure.

Producing irregular phonation by machine: In the literature, various methods were presented that aim to artificially create irregularly phonated speech. Two different approaches can be distinguished: applying formant synthesis (copy synthesis) and using waveform manipulation (that usually means a substantial increase of jitter). On the one hand, copy synthesis allows one to create many phonation types, both normal and pathological [14-16], but appropriately setting the large number of time-varying synthesis parameters requires expertese and sometimes more than a day's work [16]. On the other hand, waveform manipulation techniques [17-19] ignore that irregular phonation has a number of acoustic characteristics besides jitter that may also play a role in its perception. Another problem with these methods is that very different waveforms can yield the same jitter value (e.g. randomly perturbing fundamental periods versus systematically varying them).

The occurrence of irregular phonation in Hungarian speech: I am not aware of earlier studies investigating the occurrence of irregular phonation in Hungarian. Studies that mention irregular phonation (e.g. [9;12]) did not systematically examine its occurrence. However it is often highlighted as a problem in computer speech analysis [10;11;20].

The occurrence of irregular phonation and the speaker: It was observed in several languages that the occurrence rate of irregular vocal fold vibration can differ among speakers [5;6;21]. However these studies were not aimed at examining the speaker-dependence of this characteristic, thus they employed a small number of speakers and measured differences only among speakers, not within speakers (i.e. if speakers use irregular phonation with a similar frequency in repeated recordings). Although there are several results on the perception of irregular phonation, listeners' memories of familiar speakers' phonation habits have not been examined earlier.

2. Research objectives

My goal was to develop methods that make it possible to exploit the linguistic and non-linguistic information carried by irregular phonation in speech technologies.

Accordingly, my first goal was to create a phonation type classifier that is capable of classifying vowels as regular or irregular with higher accuracy than earlier systems (with a higher hit rate and lower false alarm rate). Note that, besides vowels, it is also possible to produce voiced consonants with irregular phonation but, for most practical applications, it is sufficient to have accurate phonation type information about the

vowels of an utterance. For example, in order to make inferences about the prosodic structure of the utterance, about the emotional state or the identity of the speaker, a syllable-level resolution (that is provided by the decisions made about each vowel) is usually satisfactory.

My second goal was to propose a procedure that can produce speech with irregular phonation by machine, i.e. which can transform a selected region of regular voice in a way that the resulting speech approaches irregular phonation both perceptually and acoustically. Note that the reverse transformation (turning irregularly phonated speech into regular) is less relevant in terms of practical applications, as the majority of our spoken utterances is produced with regular phonation.

My third goal was to investigate the idiosyncratic nature of the occurrence rate of irregular phonation. On the one hand, this meant the quantitative research of the systematic differences across speakers in a speech corpus. On the other hand, this goal refers to the experimental examination of the question whether listeners remember familiar speakers' phonatory habits. A further goal is the analysis of the general occurrence of irregular phonation in Hungarian speech. These goals point to potential applications of the above two machine procedures (for classification and for transformation of phonation types).

The majority of the work corresponding to the above goals was carried out for American English but the author focused on methods that may also be applicable to Hungarian.

3. Methodology

The research methods applied are widespread in speech technology and in related disciplines.

For training and testing the phonation type classifier, I employed the widely used TIMIT speech corpus [22]. This choice makes it possible to compare and reproduce my results. Accordingly, the accuracy of the classifier was compared with another phonation type classifier that used the same test and training set. The classifier made the phonation type decisions by means of a Support Vector Machine (SVM) that I implemented using a publicly available Matlab toolbox (OSU SVM[2]).

The phonation type transformation method was evaluated both acoustically and perceptually by generally accepted methods. In the acoustic analysis, I measured the parameters of the glottal source by the relative amplitudes of the harmonics, using the method of Holmberg et al. and Hanson [23;24]. This method allows one to measure source parameters without the need for inverse filtering (that is usually inaccurate for irregularly phonated speech). In the perceptual test, subjects rated the speech stimuli using the widespread 5-point scale. In both of these evaluations, I compared the results to irregular phonation in natural speech.

[2] http://sourceforge.net/projects/svm

4. New results

4.1. Thesis group I: Method for automatic classification of vowels produced with regular and irregular phonation

A number of phonation type (regular vs. irregular) classifiers have been proposed in the literature. There is little or no overlap in the acoustic cue sets that these classifiers use. The different cues probably capture different aspects of irregular phonation, thus one can expect a performance gain by integrating the cues described in these papers into one phonation type classifier. This thesis group proposes such a system.

The goal of the system presented here is to classify vowels into regular and irregular classes. For training and testing, I employed the recordings of speakers in dialect region 1 and 2 (New England and Northern, respectively) from the TIMIT corpus [22]. For these recordings, phonation type labels (i.e. the desired output of a phonation type classifier) were available[3] that made it possible to apply supervised learning methods. I divided the sound material into training and test sets according to the recommendation of the creators of TIMIT.

Thesis I.1: [C2] *I applied a novel analysis method (ROC curves, that are widespread in other domains) to evaluate acoustic cues used in phonation type classifiers. ROC analysis allows the assessment and quantification of the regular-irregular separation capability of each cue with no underlying assumptions about the cue value distributions and independently of the threshold value.*
I employed ROC analysis to evaluate six acoustic cues presented in the literature [4;7]:

- *Fundamental frequency (F0)*
- *Normalized RMS intensity (NRMS)*
- *Smoothed energy difference (SED)*
- *Power peak falling and rising (PWP)*
- *Shift-difference amplitude (SD)*
- *Intraframe periodicity (IFP)*

Based on these, I created five acoustic cues (by algorithmic changes and by fine-tuning the constants used during the calculation of the cue value), that are capable of distinguishing the two phonation types with higher accuracy than the original versions of these cues. Average classification performance was measured by the area under the ROC curve.

I evaluated the classification performance of acoustic cues used by earlier phonation type classifiers by means of the well-known Receiver Operating Characteristic (ROC) curves that give a more comprehensive assessment than other methods applied earlier (e.g. histograms, box-and-whisker diagrams and comparisons of means). The Area Under Curve (AUC) shows the classification capabilities of each cue independently of the actual threshold value and it is a quantitative performance measure that does not make any prior assumptions about the distribution of the cue values [25].

[3] The phonation type labels were created and shared with me by Kushan Surana, for which I am grateful.

I re-implemented all the four acoustic cues published by Surana [4] (F0, NRMS, SED and SD) and two of the three cues described by Ishi et al. [7] (PWP and IFP; the third cue serves to separate voiced and voiceless regions of speech and thus it is not relevant for my goal). Then, I improved the regular-irregular separation capability of these cues in two ways. On the one hand, I optimized the constants used by the algorithms of the cues by systematic tests. On the other hand, I made changes in the algorithms and examined potential alternatives. Besides these modifications, for the frame-based cues, the framing procedure was changed to use overlapping frames (30 ms long with 5 ms steps) in order to increase the time-resolution. The basic idea behind the original cues and the changes made are briefly summarized below (detailed descriptions can be found in the dissertation):

Fundamental frequency (F0): For irregularly phonated speech, a traditional autocorrelation-based pitch tracker is expected either to return F0 values lower than for regular speech or to consider that region of speech unvoiced (denoted by a return value of 0 Hz) [4].
Due to the finite frame length, the normalized autocorrelation function has a linearly decaying envelope that reaches zero at the maximum lag. For this reason, I used the unbiased autocorrelation function instead, so that the peaks corresponding to a low fundamental frequency are not excluded from the analysis.
The value of the voicing threshold was set to 0.35, because the systematic tests performed over the range 0.25-0.50 showed that this yields higher performance than the original value of 0.46. Further, one of the criterions for selecting the autocorrelation peak corresponding to F0 was removed, as it usually resulted in halving errors. As a result of these changes, the area under the ROC curve increased from 0.87 to 0.93.

Normalized RMS intensity (NRMS): The RMS intensity, normalized to the entire sentence, is expected to be lower for irregular voice than for regular voice (in irregularly phonated speech, there are generally fewer glottal pulses in a given time frame than in regularly phonated speech) [4].
However, normalizing for the entire sentence can be misleading as there may be substantial intensity changes along the sentence. This is why instead of the entire sentence, I rather normalize to the vowel and its environment. Tests showed that normalizing to the vowels and its ±25 ms environment slightly increases the area under the ROC curve (from 0.84 to 0.86).

Smoothed energy difference (SED): This cue attempts to capture the rapid energy transitions in irregular phonation that are due to the wider spacing of the glottal impulses. In order to do that, it calculates an energy contour with a fine time-resolution, smoothes it by two different window lengths (6 ms and 16 ms), and takes the maximum difference of the two smoothed energy contours [4].
Tests examining the effect of various window size combinations revealed that using a 2 ms and a 4 ms smoothing window can improve the separation of regular and irregular tokens. As another change, instead of taking the maximum, I calculate the absolute maximum at the last step of the algorithm. After these changes, the area under the ROC curve increased by 0.08 (from 0.74 to 0.82).

Power peak rising and falling (PWP): Similarly to SED, this cue is based on the rapid energy transitions around irregular glottal impulses. The peaks of a “very-short-

term power contour" are picked, and then a measure of the rate of power rise and fall (before and after the peak) is computed [7].
The two constants employed by the algorithm (the distance of the samples considered before and after the sample in question for finding the peaks in the "very short term power contour", and the length of the environment used to calculate the degree of power fall and rise) were optimized by systematic tests (resulting in values of 4 instead of 3, and 4 instead of 5, respectively) and, as a result, the performance of the cue improved: the area under the ROC curve rised from 0.79 to 0.85.

Shift-difference amplitude (SD): This cue estimates the harmonic-to-noise ratio (HNR) in the speech signal. It calculates the minima (sample-by-sample) of squared difference signals between the input and its copies shifted with a range of lags, it normalizes them and takes their mean over time [4].
The "aperiodicity measure", published by Kochanski et al. [26], is an alternative method to calculate the shift-difference amplitude (the original SD cue was inspired by this measure). The output of this algorithm is the harmonic-to-noise ratio as the function of time. The mean of this function was tested as an alternative to the original SD cue. This alternative cue increases the area under the ROC curve from 0.75 to 0.88.

Intraframe periodicity (IFP): As the shift-difference amplitude, IFP also aims to measure harmonic-to-noise ratio in speech. The first prominent peak (corresponding to a positive lag) in the unbiased autocorrelation function is picked. Then, the autocorrelation values at integer multiples of this peak lag are obtained and their minimum is selected as the measure of intraframe periodicity [7].
The IFP cue provides an excellent separation between regular and irregular vowels in its original form and thus I did not aim to suggest improvements for this cue.

For evaluating the cues described, the cue values for all the vowels in the training set were computed, and then, using the phonation type labels, the ROC curves were plotted and the AUC values were calculated. The change in the area under the ROC curve was used as a quantitative measure of the effect of the changes: if the area increased, the average classification performance of the cue improved (Figure 2, solid lines).

It can be concluded that, to some degree, each cue is capable of separating vowels produced with regular and irregular phonation (the majority of each ROC curve is above the diagonal representing the chance level, and accordingly all the AUC values are larger than 0.5). The best classification performance was achieved by IFP, followed by F0. A short section of the ROC curve belonging to the SED cue runs below the diagonal. This suggests that a classifier based on the SED cue and using a threshold corresponding to a point in that section would make mainly incorrect decisions. However the majority of the SED ROC curve is above the diagonal and therefore, in most cases, it can be a useful cue.

As it can be seen on the figure, the ROC curves of the modified cues lie above the curves corresponding to the original cues. This suggests that the overall classification performance increased in all five cases. The increases in the areas under the ROC curves show that the suggested changes yielded a performance gain.

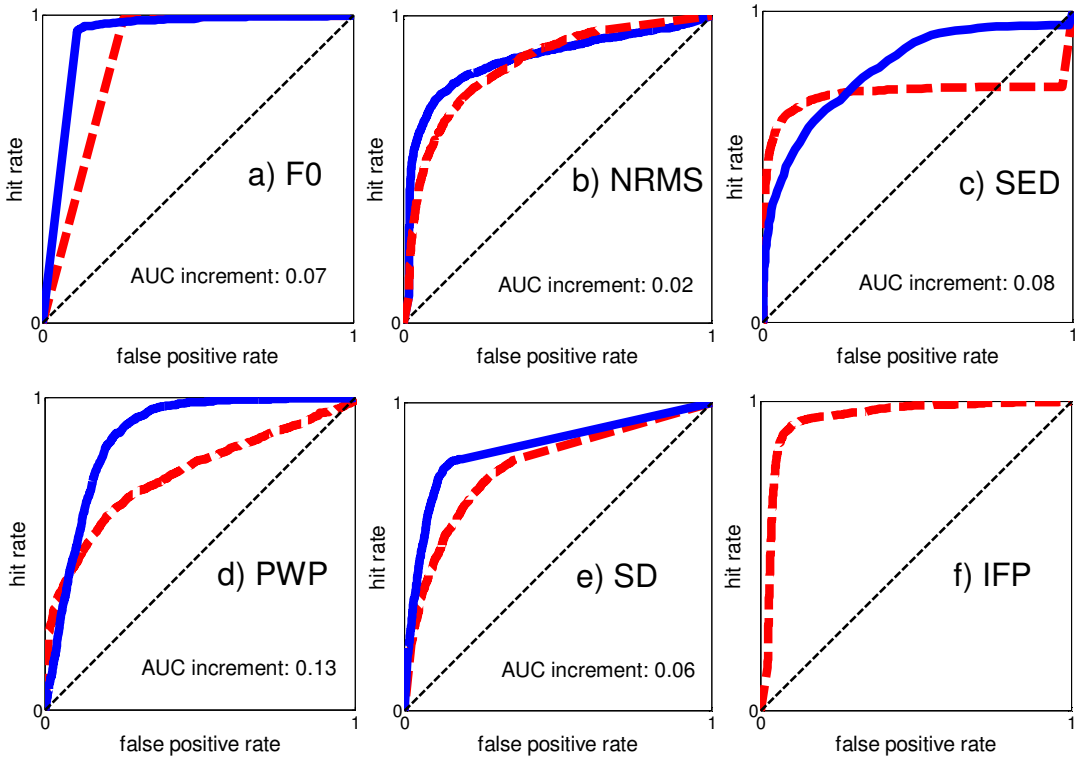

Figure 2: The ROC curves of the six acoustic cues (calculated on the training set). The dashed lines correspond to the original versions of these cues, while the solid lines belong to the modified algorithms.

Thesis I.2: [C2] *I designed and implemented an automatic phonation type classifier that can classify vowels as regular or irregular with higher accuracy than an earlier, comparable system [4]. The classifier integrates the acoustic cues described and improved in Thesis I.1 and, based on these cue values, makes a phonation type decision by means of an SVM. Thus I showed that by combining and improving the acoustic cues from several earlier systems, a substantial performance gain can be achieved.*

The SVM-based phonation type classifier uses the six acoustic cues (in five cases, the improved versions) as features. The training of an SVM with an RBF (Radial Basis Function) kernel is contolled by two parameters: C, the cost of incorrect classifications, and γ, a property of the Gaussian kernel. I ran a grid search on a wide range of values for these two parameters. A subset of the training set (1380 regular and 1380 irregular vowels) was used for the grid search. For each parameter combination, both a 3-fold and a 10-fold cross-validation was performed and the averages of the two hit rates and two false alarm rates was calculated. The two averaged performance measures were then combined by an equal weight to obtain the accuracies for each parameter setting. The highest accuracy was achieved by using C=0.0313 and γ=0.0313.

The SVM was trained, using the above parameter values, on a balanced subset of the training set (containing the same number of regular and irregular items). This

balanced subset contained all the 1403 regular vowels and the same number of randomly chosen irregular vowels.

On the test set, the proposed phonation type classifier achieved a 98.85% hit rate and a 3.47% false alarm rate. These results can be compared with those of Surana [4] as that system also makes regular/irregular decisions on vowels. Surana reported a hit rate of 91.25% and a false alarm rate of 4.98% on the same test set. The system described in this thesis thus achieved a 7.60 percentage points higher hit rate and a 1.51 percentage points lower false alarm rate than this earlier, comparable classifier. The performance achieved by the classifier is likely to be suitable for most practical applications.

4.2. Thesis group II: Procedure for transforming regular voice into irregular voice

This thesis group introduces a procedure that is capable of transforming speech produced with regular phonation to irregular voice. It introduces irregular pitch periods into a modal speech signal by scaling the amplitude of the individual cycles. The transformed speech is evaluated both perceptually and acoustically, using natural (regular and irregular) speech for comparison.

Thesis II.1: [C3, C4, J1] *I developed a method to transform regular voice into irregular voice by scaling the amplitudes of the individual fundamental periods (including scaling them to zero). The procedure of 'stylized pulse pattern copying' was proposed for setting the scaling factors semi-automatically.*

First the fundamental periods are approximately separated by pitch-synchronous windowing, then their amplitudes are individually scaled by scaling factors (often scaled to zero, i.e. removing cycles from the signal), and finally overlapped and added. The input is the speech waveform ($x[n]$) with markers for the approximate times of glottal excitations (*pitch marks*; p_i, $1 \le i \le N$) in the region to be transformed. As a first step, a rough approximation of the fundamental periods is extracted by applying a Hanning-window ($w_i[n]$, $1 \le i \le N$) in the vicinity of each pitch mark. The peak of the window is positioned on the pitch mark and it spans from the previous to the next pitch mark (thus it covers two fundamental periods and it may be asymmetric). This windowing procedure is the same as the analysis stage of the PSOLA (Pitch Synchronous OverLap-and-Add) algorithm [27] and it extracts rough estimates of the individual glottal cycles into separate waveforms. The samples in each of these waveforms are then multiplied by a hand-selected scaling factor (s_i) and overlapped-and-added to re-synthesize the signal:

$$\hat{x}[n] = \sum_{i=1}^{N} \left(s_i x[n] w_i[n] \right)$$

where $\hat{x}[n]$ is the output speech signal. The scaling factors can either boost (s_i>1), attenuate (s_i<1), remove (s_i=0) or leave unmodified (s_i=1) the fundamental periods. See Figure 3b for an example of a transformed speech waveform.

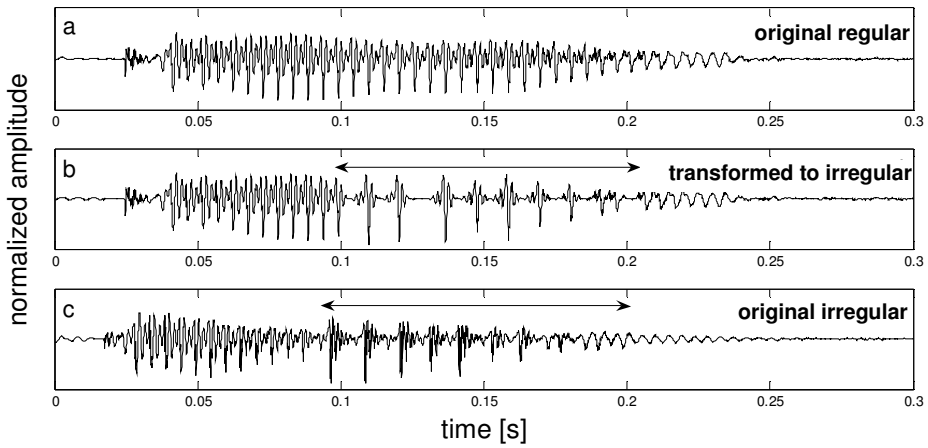

Figure 3: A speech recording with a regular ending (a) and its transformed version (b). An originally-irregular recording is also shown (c). Horizontal arrows mark irregular regions.

Attenuating or zeroing an impulse response also scales down the background noise present during that fundamental period (e.g. when several consecutive cycles are removed from a recording, the lack of audible noise might decrease the perceived naturalness). In order to avoid this problem, background noise ($b[n]$) can be added to attenuated and zeroed impulse responses. The noise can be windowed out from e.g. the end of the recording and scaled by 1-s_i in order to compensate for the noise energy loss due to scaling:

$$\hat{x}[n] = \sum_{i=1}^{P} \left(s_i x[n] w_i[n] + \max(1 - s_i, 0) b[n - p_i] w_i[n] \right)$$

The scaling factors can be set semi-automatically by the method of "impulse pattern copying". This operates by modeling the desired impulse pattern (the time spacings and amplitudes of the glottal impulses) based on a region of natural speech with irregular phonation – i.e. a model recording is used, from which the impulse pattern is copied to the speech signal to be transformed. The pulse pattern is initially constructed as a vector containing the relative amplitudes of the glottal pulses in the sample irregular region. When an irregular cycle is substantially longer than a reference cycle length (e.g. two or three times or more than the reference, $T0_{ref}$, that is calculated as the mean of several preceding regular cycles), an appropriate number of zeros are inserted in the impulse pattern since, at these points, periods need to be removed from the regular recording.

Thesis II.2: [J1, C3, C4] *I designed a perceptual experiment that showed that speech transformed with the procedure of Thesis II.1 is judged by the listeners to have a roughness similar to irregular phonation and a naturalness close to natural speech.*

A major factor in determining the practical usefulness of the transformation method proposed in Thesis II.1 is its acceptability to human listeners. This was measured in a

listening test that aimed to evaluate the degree of perceived roughness and any degradation in naturalness introduced in the speech signal by the method. As the perception of irregular phonation is usually described as rough voice, measuring perceived roughness allows the assessment of whether the transformed recordings sound like irregular phonation. Apart from roughness, a significant degradation in naturalness would signal audible artifacts introduced by the method. In the evaluation, I focused on irregular phonation in utterance-final positions. I compared the results of transformed speech samples to unmanipulated natural utterances, both with regular and with irregular phonation at the end.

The perceptual experiment consisted of two tests. In one, listeners (12 native speakers of American English) rated the naturalness of the speech samples, while in the other they judged the roughness of the same samples. For both tests, responses were given on a 5-point scale.

The results are shown on Figure 4. When a natural utterance with regular utterance-final phonation was transformed to irregular voice with the proposed method the roughness ratings increased by 1.09 point ($p<0.0005$), but this transformation caused only a non-significant, 0.19 point decrease in naturalness scores ($p=0.11$, not significant). Not only did the transformation substantially increase perceived roughness, but this increase matches the roughness of natural irregularly-phonated speech: the difference in the mean ratings was only 0.11 points ($p=0.91$, n. s.). Thus, listeners perceived the increased roughness of the transformed utterances, considered them natural, and heard no difference in the degree of roughness between the originally-irregular and transformed-irregular stimuli.

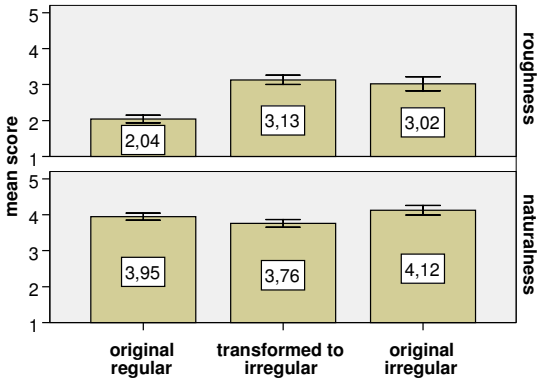

Figure 4: Mean scores from the perceptual evaluation. Vertical bars correspond to 95% confidence intervals.

Corollary: *In contrast to earlier methods, the procedure of Thesis II.1 introduces abrupt, substantial changes in the timing and amplitude of the fundamental periods (e.g. doubles the period length by removing a cycle). According to the results of the perceptual evaluation, these abrupt, substantial changes in the amplitude and duration of glottal cycles can elicit the roughness percept similar to natural irregular phonation.*

Thesis II.3: [C3] *By adopting a well-known measurement procedure, I showed that the transformation method of Thesis II.1 changes a number of acoustic parameters (spectral tilt, open quotient, first formant bandwidth) towards the values characteristic of irregular phonation.*

Irregular phonation, compared to regular, is usually characterized by five acoustic properties (for detailed explanations, see [6;14]): (1) F0 is lower and jitter is higher; (2) the overall intensity is lower; (3) the open quotient (OQ) is lower; (4) the bandwidth of the first formant (B1) is increased; and (5) the spectral tilt (TL) is lower. It is clear that, because it involves the removal of individual pitch periods, the transformation method substantially increases and perturbs the spacing of the glottal pulses, and also decreases the overall intensity level.

In order to examine the effect of the transformation on parameters (3)-(5), I conducted a set of acoustic measurements. The measurements were performed in the frequency domain, using the method of [23] (with corrections described in [24]): OQ was approximated by the first harmonic amplitude relative to the second harmonic amplitude (i.e. by H1*-H2* in dBs); TL by the first harmonic amplitude relative to the third formant amplitude (H1-A3*) and B1 by the first harmonic amplitude relative to the first formant amplitude (H1*-A1). These measurements were conducted on the stimuli used in the perceptual evaluation in Thesis II.2, i.e. original regular and irregular recordings and the former ones transformed to irregular. Details of the measurements can be found in the dissertation.

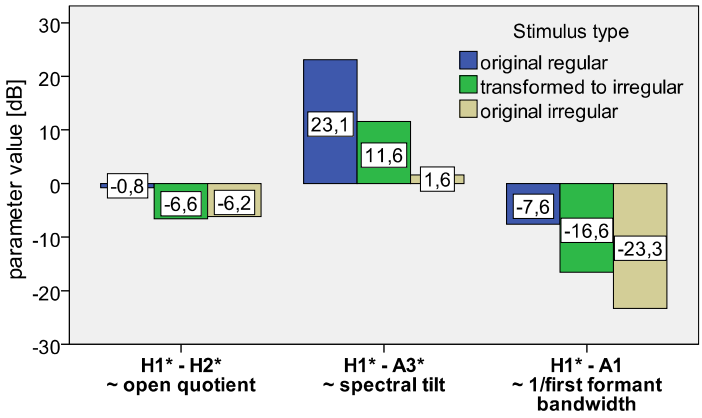

Figure 5: Averages of the three measured acoustic parameters

Results are displayed in Figure 5. H1*-H2* of the transformed recordings was significantly lower than that of the original regular speech samples ($p<0.0005$) and was approximately the same as the mean value of the original irregulars ($p=0.97$, n.s.). Thus, in terms of open quotient, the transformed utterances closely matched the values for natural irregular phonation. H1*-A3* (and correspondingly spectral tilt) also decreased significantly due to the transformation ($p=0.001$), moving toward the mean value for irregular speech but still being different from it ($p=0.033$). H1*-A1 also showed significantly lower values after the transformation than before ($p<0.0005$), that is, B1 increased (note that B1 is inversely proportional to H1*-A1). Although the mean H1*-A1 for transformed speech is closer to the average value of

that parameter for original irregular tokens than for natural regulars, it is still significantly higher ($p=0.001$).

It can be concluded that, besides lowering F0 and overall energy and introducing jitter, the transformation method of Thesis II.1 also reproduces some additional acoustical correlates of perceived roughness, such as a decreased open quotient and spectral tilt, and an increased B1. This represents a potential improvement over earlier manipulation methods that aimed to increase the roughness of a speech signal by focusing on jitter and shimmer.

4.3. Thesis group III: The occurrence of irregular phonation and the speaker

This thesis group presents, on the one hand, the results of a study analyzing both the general and the speaker-dependent occurrence of irregular voice in Hungarian. On the other hand, it presents two experiments on English that investigated whether subjects remember speakers' usual utterance-final phonation type (for familiar voices) and whether they can store this information in memory after a short training (for previously unknown voices).

Thesis III.1: [B1] *I designed and performed an experiment showing that irregular voice occurs in normal Hungarian speech and that the occurrence rate of sentence-final irregular phonation can be substantially different between speakers, while exhibiting only smaller variation within a speaker.*

12 native speakers of Hungarian (2 females, 10 males, aged 23-33, except one male speaker, who is 64 years old) were recorded while reading 11 Hungarian sentences (7 declarative, 3 interrogative and 1 exclamation). After reading all the 11 sentences, the sentence set was read two more times.

Instances of irregular phonation were labeled by hand. In order to improve consistency, consensus labeling was applied: besides the author, another expert labeled the entire speech corpus independently and then these parallel label files were compared. The differences were discussed and a merged label file was created (the method of merging is described in the dissertation).

Irregular phonation was found in 9% of all the syllables. This ratio is remarkably high, especially if one considers that irregular phonation has been mostly ignored so far. This result shows however that irregular phonation is not a rare, negligible phenomenon in Hungarian speech: in the speech corpus examined, on average, every 11th syllable was uttered using this phonation type.

A label was considered sentence-final if it reached into the last syllable of a sentence or into the one before the last syllable. According to Figure 6, the occurrence rate of irregular phonation shows large differences across the speakers. 3 speakers frequently used this phonation type, while another 5 speakers seldom did it. The occurrence rate of irregular phonation for a given speaker was similar across the three repetitions. In case of 7 speakers, the three repetitions contained roughly the same number of

syllables produced with irregular phonation, and for 5 speakers, one can see moderate differences.

Results showed that, in agreement with earlier informal observations, the frequency of irregular phonation is probably characteristic to the speaker, especially at the ends of sentences. When comparing the three repetitions of the same texts, subjects uttered roughly the same percentage of syllables with irregular voice.

Corollary: *It is likely that, by appropriately handling this phonation type in speech technologies for Hungarian, a considerable performance gain can be achieved. Further, the frequency of irregular phonation can be a useful cue (in addition to the cue set already available) in machine speaker identification and speaker verification systems, as well as in forensic voice analysis.*

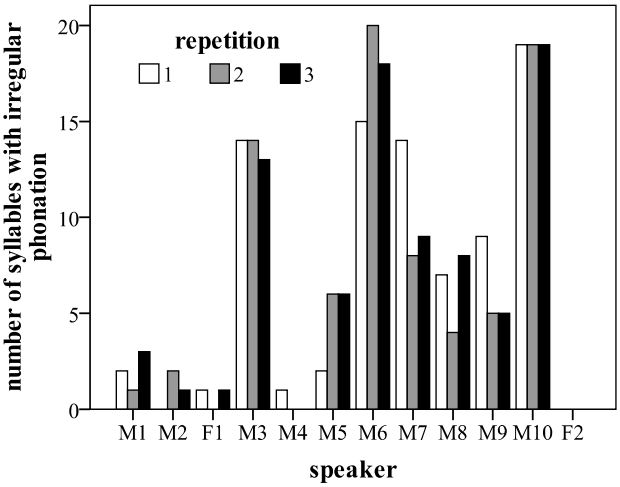

Figure 6: Number of occurrences of irregular phonation in sentence-final position (in the last syllable and/or in the syllable before the last one in the sentence) by speaker and by repetition.[4]

Thesis III.2: [J1, B3, C5, C8] *With an experiment in English, I showed that listeners remember familiar speakers' utterance-final phonation habits (the likelihood of irregular phonation).*

Based on the results reported in Thesis III.1, it appears that speakers differ in their likelihood of producing irregular pitch periods in utterance-final syllables. In this thesis, I examined whether listeners remember this variation in a familiar speaker's voice and can access this information.

The 9 listeners personally knew the speakers – one male and one female whose speech frequently exhibited utterance-final irregular pitch periods and one male and one female whose speech seldom did. Four short words and phrases uttered by each of the four speakers were used (all of these utterances exhibited the speakers' habitual utterance-final phonation type).

[4] There was no sentence-final irregular phonation found in speaker F2's utterances.

I created three manipulated versions each of the 16 recordings: one with a modified phonation type, one with a modified mean F0, and one with both of these modifications. The phonation type modification was carried out by the procedure described in Thesis group II. During the experiment, listeners were presented with minimal pairs (whose naturalness was tested in a separate experiment): one member of the pair was an unmanipulated original recording and the other one was a manipulated version of that recording, in random order. Thus the pairs differed (a) only in utterance-final phonation type, (b) only in mean F0, or (c) in both parameters. The task of the listeners was to decide which member of the pair belongs to the familiar speaker.

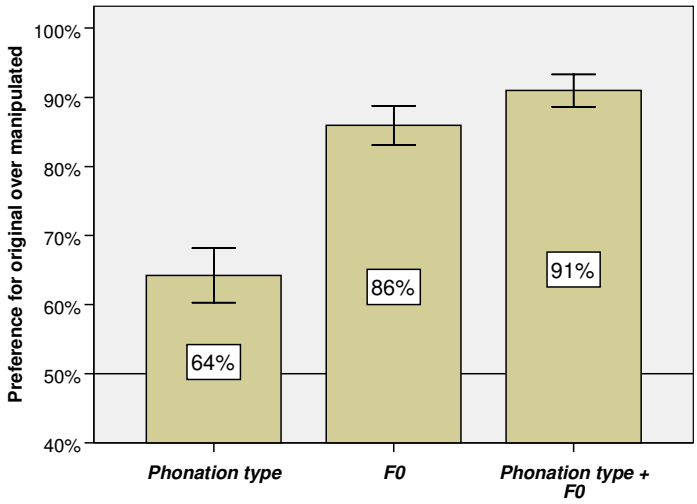

Figure 7: Preference for the stimuli with the speakers' usual phonation type, mean F0, or both over their manipulated versions. The horizontal line corresponds to the 50% chance level and error bars represent 95% confidence intervals.

The main results can be seen in Figure 7, which shows the proportion of correct responses (where 'correct' means that the original unmanipulated speech sample was preferred over a manipulated one) for the three experimental conditions. The *Phonation type* condition served to test the hypothesis that the typical pattern of utterance-final irregularity is a feature that listeners remember about the speaker. Listeners preferred the speakers' habitual phonation type over tokens with changed phonation type 64% of the time, that is significantly higher than the 50% chance level ($t(575)=7.122$; $p<0.0005$). Although there was considerable variation across speakers and listeners in the effectiveness of this memory recall, utterance-final phonation type had a significant effect in a substantial number of cases (for all four speakers, and for 5 out of 9 listeners).

For the cases where the F0 contour was shifted up or down for the transformed member of the pair, the preference rate for the unmanipulated member was 86%. This is consistent with results in the literature showing that mean F0 is a robust cue to speaker identity [28]. Manipulating phonation type showed little interaction with changing F0: the significant increase in the rate of correct responses for the *Phonation*

type+F0 condition compared to the *F0* condition (t(1150)=2.680; p=0.007) suggests that, even when a voice characteristic as effective as F0 is available, the appropriate pattern of utterance-final irregularity still makes it slightly easier for listeners to tell which of two speech samples was produced by the target speaker.

Corollary: *When recognizing familiar voices, this phonation type information is available for the cognitive process performing the recognition and can be used as a cue.*

Thesis III.3: [C6] *I showed that listeners encode in memory the characteristic occurrence rate of utterance-final irregular phonation of an unfamiliar speaker during brief perceptual learning.*

While in the experiment of Thesis III.2, listeners were familiar with the speakers' voices, in this experiment I recruited unfamiliar listeners (12 persons). This experiment tested whether speakers' characteristic occurrence rate of utterance-final irregular phonation can be learned in a brief training.

The experiment consisted of an initial training phase (during which listeners watched a video recording of the four speakers reading a short story), and three test phases. In each test phase, there were two tests: the first one measuring how accurately listeners can recognize the speakers' by voice (allowing to track listeners' progress in learning the voices). The second one was similar to the paired comparisons test of Thesis III.2.

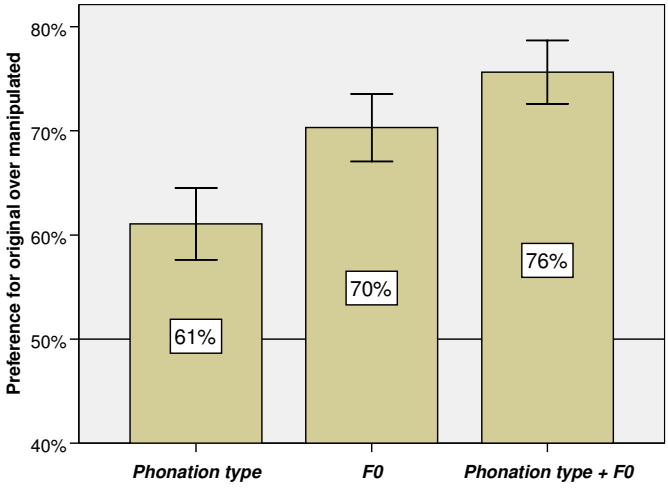

Figure 8: Preference for the stimuli with the speakers' usual phonation type, mean F0, or both over their manipulated versions after training. The horizontal line corresponds to the 50% chance level and error bars represent 95% confidence intervals.

According to the results of the speaker recognition tests, in the first phase listeners were not familiar enough yet with the speakers' voices. Consequently, I considered the first phase to be practice, and excluded the paired comparison responses of this phase from further data analysis. Figure 8 shows the proportion of correct responses for the other two paired comparisons tests that displays a similar pattern to that seen in

Thesis III.2. When utterance-final phonation type was manipulated, 61% of the responses were correct, i.e. tokens with the original phonation type were preferred over tokens with changed phonation type 61% of the time (significantly higher than the 50% chance level, $t = 5.328$; $p < 0.0005$). This is possible only if listeners store in memory the speakers' phonatory habits in many cases.

The results showed that brief perceptual learning without feedback is sufficient for listeners to encode information about the talker's characteristic utterance-final phonation type in memory.

Corollary (Theses III.2 and III.3): *Based on these results, related speech technologies, such as personalized speech synthesis and voice conversion, may benefit from using such intermittent acoustic patterns.*

5. Applicability of the results

In the Introduction, I briefly described the three potential roles of irregular phonation in speech communication (carrying information about the linguistic content, about the emotional state of the speaker and about the identity of the speaker). Accordingly, the phonation type classifier introduced in Thesis group I can be applied in three areas of speech technology. First, it can provide supplementary information in speech recognizers: for example, it can contribute to detecting intonational phrase boundaries and thus making it possible to splice the input into smaller units (that can be processed more efficiently by the speech recognizer). Second, systems classifying the emotional state of the speaker based on his/her speech can benefit from using phonation type information as an acoustic cue. Third, based on the results of Thesis III.1, an automatic speaker identification system can employ the occurrence rate of utterance-final irregular phonation in separating different speakers' voices.

The potential applications of Thesis group II can also be presented along the three communicative functions of phonation type. First, in text-to-speech systems, the transformation method can make the generated prosody more natural-sounding if irregularities are introduced in the speech signal at appropriate locations (e.g. at intonational phrase boundaries). Second, transforming the voice to irregular can help in expressing certain emotions in synthetic speech. Third, based on results presented in Theses III.2 and III.3, handling irregular phonation can contribute to personalized speech synthesis (that is adapted to a specific talker's voice). In order to allow fast and convenient application of the transformation method of Thesis II.1, I designed and implemented a software tool. The graphical user interface is the work of Nicolas Audibert (GIPSA-lab, Grenoble). The program is freely available on the internet [5].

Besides their relevance in phonetics and psycholinguistics, results of Thesis group III point out that the appropriate modeling of irregular phonation (e.g. by means of the methods described in Thesis groups I and II) can contribute to improving the quality of several speech technologies, both in Hungarian and in English.

[5] http://www.bohm.hu/glottalizer.html

Publications

Journal articles

[J1] Tamás Bőhm, Stefanie Shattuck-Hufnagel: Do listeners store in memory a speaker's habitual utterance-final phonation type? *Phonetica*, volume 66, issue 3, pp. 150-168, 2009.

[J2] Bőhm Tamás: Beszélőfelismerés – neurológiai háttér és pszichológiai modellek. *Magyar Pszichológiai Szemle* 62 (4), 2007. december, pp. 541-563.

[J3] Tamás Bőhm, Géza Németh: Algorithm for formant tracking, modification and synthesis, *Infocommunications Journal*, vol. LXII., 2007/1 Selected papers, pp. 15-20.

> [J3b] Bőhm Tamás, Németh Géza: Algoritmus formánsok követésére, módosítására és szintézisére, *Híradástechnika*, LXI. évfolyam, 2006/8, pp. 11-16.

[J4] Németh Géza, Olaszy Gábor, Bőhm Tamás, Ugron Zoltán: Szöveges adatbázis tervezése rendszerüzenet generátorhoz, *Híradástechnika*, LXI. évfolyam, 2006/3, pp. 38-42.

[J5] András Nagy, Péter Pesti, Géza Németh, Tamás Bőhm: Design issues of a corpus-based speech synthesizer. *Infocommunications Journal*, vol. LX., 2005/6 Selected papers, pp. 6-12.

> [J5b] Nagy András, Pesti Péter, Németh Géza, Bőhm Tamás: Korpusz-alapú beszédszintézis rendszerek megvalósítási kérdései. *Híradástechnika*, LX. évf., 2005/1, pp. 18-24.

[J6] Tamás Varga, Péter Benkő, Tamás Bőhm, Attila Eschwig-Hajts: Fluid simulation in telecommunication networks. *Infocommunications Journal*, vol. LVII, 2002/7 Selected papers, pp. 41-45.

Chapters in edited books

[B1] Bőhm Tamás, Ujváry István: Irreguláris fonáció előfordulása magyar beszédben, mint egyéni hangjellemző. In: Gósy Mária (szerk): *Beszédkutatás*, MTA Nyelvtudományi Intézet, Budapest, 2008. pp. 108-120.

[B2] Bőhm Tamás, Olaszy Gábor: A magyar [v] hang szerkezetének és zörejességének fonetikai vizsgálata. In: Gósy Mária (szerk): *Beszédkutatás*, MTA Nyelvtudományi Intézet, Budapest, 2007. pp. 19-34.

[B3] Bőhm Tamás: A glottalizáció szerepe a beszélő személy felismerésében. In: Gósy Mária (szerk): *Beszédkutatás*, MTA Nyelvtudományi Intézet, Budapest, 2006. pp. 197-207.

Papers in conference proceedings

[C1] Tamás Gábor Csapó, Zsuzsanna Bárkányi, Tekla Etelka Gráczi, Tamás Bőhm, Steven M. Lulich: Relation of formants and subglottal resonances in Hungarian vowels, *Proc. Interspeech 2009*, September 6-10, 2009, Brighton, United Kingdom. pp. 484-487.

[C2] Tamás Bőhm, Zoltán Both, Géza Németh: Automatic classification of regular vs. irregular phonation types, *Proc. NOLISP 2009*, June 25-27, 2009, Vic, Spain. pp. 53-61.

[C3] Tamás Bőhm, Nicolas Audibert, Stefanie Shattuck-Hufnagel, Géza Németh, Véronique Aubergé: Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles, *Proc. Acoustics 2008*, Paris. pp. 6141-6146.

[C4] Bőhm Tamás, Németh Géza: Egy egyszerű módszer modális beszéd glottalizálttá alakítására, *V. Magyar Számítógépes Nyelvészeti Konferencia*, 2007. december 6-7., Szeged, pp. 267-270.

[C5] Tamás Bőhm, Stefanie Shattuck-Hufnagel: Utterance-final glottalization as a cue for familiar speaker recognition, *Proc. Interspeech 2007*, August 27-31, 2007, Antwerp, Belgium, pp. 2657-2660.

[C6] Tamás Bőhm, Stefanie Shattuck-Hufnagel: Listeners recognize speakers' habitual utterance-final voice quality, *Proc. International Workshop on Paralinguistic Speech*, August 3, 2007, Saarbrücken, Germany, pp. 29-34.

[C7] Géza Németh, Géza Kiss, Tamás Bőhm: An algorithm for formant tracking, modification and re-synthesis of speech. *Trends in Speech Technology*, May 13-14, 2005, Cluj-Napoca, pp. 59-66.

Conference abstracts

[C8] Tamás Bőhm: Is utterance-final glottalization a cue for speaker recognition by humans? *151st Meeting of the Acoustical Society of America*, June 5-9, 2006, Providence. Journal of the Acoustical Society of America, volume 119, issue 5, May 2006, pp. 3244-3245.

[C9] Tamás Bőhm, Géza Németh, Géza Kiss: A visual tool for the demonstration of formants in speech, *International Workshop in Phonetics Dedicated to the Memory of Farkas Kempelen*, March 11-13, 2004, Budapest.

References

[1] Fék, M., Pesti, P., Németh, G., Zainkó, Cs., "Generációváltás a beszédszintézisben," *Híradástechnika*, vol. LXI, no. 3, pp. 21-30, 2006.

[2] Tóth, B., Németh, G., "Rejtett Markov-modell alapú mesterséges beszédkeltés magyar nyelven," *Híradástechnika*, vol. LXIII, pp. 2-6, 2008.

[3] Szarvas, M., Fegyó, T., Mihajlik, P., Tatai, P., "Automatic Recognition of Hungarian: Theory and Practice," *International Journal of Speech Technology*, vol. 3, no. 3-4, pp. 237-251, 2000.

[4] K. Surana, "Classification of vocal fold vibration as regular or irregular in normal, voiced speech." M.Eng. thesis, MIT, 2006.

[5] Redi, L., Shattuck-Hufnagel, S., "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, no. 4, pp. 407-429, 2001.

[6] Henton, C., Bladon, A., "Creak as a sociophonetic marker," in *Language, speech and mind*. Hyman, L. M., Li, C. N., Eds. London: Routledge, pp. 3-29, 1987.

[7] Ishi, C. T., Sakakibara, K. I., Ishiguro, H., Hagita, N., "A Method for Automatic Detection of Vocal Fry," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 47-56, 2008.

[8] Gósy, M., *Fonetika, a beszéd tudománya*. Budapest: Osiris Kiadó, 2004.

[9] Fónagy, I., Magdics, K., *A magyar beszéd dallama*. Budapest: Akadémiai Kiadó, 1967.

[10] A. Markó, "A spontán beszéd néhány szupraszegmentális jellegzetessége." Ph.D. dissertation, ELTE, 2005.

[11] Tóth, L., Kocsor, A., "A Magyar Telefonbeszéd-adatbázis (MTBA) kézi feldolgozásának tapasztalatai," in *Beszédkutatás 2003*. Gósy, M., Ed. Budapest: MTA Nyelvtudományi Intézet, pp. 134-146, 2003.

[12] Elekfi, L., "Hanglejtés," in *Nyelvművelő Kézikönyv,* Second ed. Grétsy, L., Kovalovszky, M., Eds. Budapest: Akadémiai Kiadó, p. 774, 1983.

[13] Hirano, M., *Clinical examination of voice*. Vienna: Springer, 1981.

[14] Klatt, D. H., Klatt, L. C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820-857, Feb. 1990.

[15] Childers, D. G., Lee, C. K., "Vocal quality factors: analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394-2410, Nov. 1991.

[16] Bangayan, P., Long, C., Alwan, A. A., Kreiman, J., Gerratt, B. R., "Analysis by synthesis of pathological voices using the Klatt synthesizer," *Speech Communication*, vol. 22, no. 4, pp. 343-368, Sept. 1997.

[17] McCree, A. V., Barnwell, T. P., III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 4, pp. 242-250, 1995.

[18] Verma, A., Kumar, A., "Introducing Roughness in Individuality Transformation through Jitter Modeling and Modification," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05),* vol. 1, pp. 5-8, 2005.

[19] Loscos, A., Bonada, J., "Emulating rough and growl voice in the spectral domain," *7th International Conference on Digital Audio Effects (DAFx'04),* pp. 49-52, 2004.

[20] Olaszy, G., Bartalis, M., "Jelfeldolgozási és fonetikai algoritmusok kombinációja a gépi hanghatárjelölés javítására," in *Beszédkutatás 2008*. Gósy, M., Ed. Budapest: MTA Nyelvtudományi Intézet, pp. 208-220, 2008.

[21] Slifka, J., "Some physiological correlates to regular and irregular phonation at the end of an utterance," *Journal of Voice*, vol. 20, no. 2, pp. 171-186, June 2006.

[22] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L, and Zue, V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, 1993.

[23] Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. C., Goldman, S. L., "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech and Hearing Research*, vol. 38, no. 6, pp. 1212-1223, Dec. 1995.

[24] Hanson, H. M., "Glottal characteristics of female speakers: Acoustic correlates," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466-481, 1997.

[25] Fawcett, T., "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[26] Kochanski, G., Grabe, E., Coleman, J., Rosner, B., "Loudness predicts prominence: Fundamental frequency lends little," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1038-1054, 2005.

[27] Moulines, E., Charpentier, F., "Pitch-Synchronous Wave-Form Processing Techniques for Text-To-Speech Synthesis Using Diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453-467, 1990.

[28] Abberton, E., Fourcin, A. J., "Intonation and speaker identification," *Language and Speech*, vol. 21, no. 4, pp. 305-318, Oct. 1978.