Listeners Recognize Speakers' Habitual Utterance-Final Voice Quality

Tamás Bőhm [1], *Stefanie Shattuck-Hufnagel* [2]

[1] Department of Telecommunications and Media Informatics, BME, Budapest, Hungary
[2] Research Laboratory of Electronics, MIT, Cambridge, MA USA

bohm@tmit.bme.hu, stef@speech.mit.edu

Abstract

This study tests the hypothesis that, among the other paralinguistic information conveyed by voice qualities, irregular vocal fold vibration (glottalization) can serve as a cue to speaker identity. Earlier studies that have reported systematic differences across speakers in the rate and type of intermittent glottalization support this idea. Still, it remains an open question whether human listeners use this speaker-specific information when recognizing familiar voices. Results of a perceptual experiment, in which listeners were first trained in order to learn the voices of the speakers, suggest that irregular pitch periods in utterance-final regions can be a cue in the recognition of individual speaker voices.

Index Terms: speaker recognition, glottalization, creak, voice quality, learning voices

1. Introduction

In this study we investigate the contribution of intermittent glottalization to a listener's ability to recognize a familiar speaker's voice. We define glottalization as perceivably irregular vocal fold vibration (see Fig. 1a for an example of glottalization occurring at the end of an utterance). We use the criterion of 'perceivable' to ensure that minor disturbances of perfect periodicity, inherent to regular phonation, are not considered, and the criterion of 'intermittent' to include speech with limited regions of aperiodicity, rather than speech that is persistently glottalized throughout the utterance. We focus on utterance-final glottalization because it usually spans a longer time than glottalization in other positions, making it clearly distinct from non-glottalized endings.

It has been traditionally assumed that glottalization is produced by the strong adduction of the vocal folds that results in low airflow through the glottis [1] (p. 122-126). However, a recent experiment involving simultaneous acoustic and physiological measurements by Slifka [2] showed that irregular vocal fold vibration can also be produced by the abduction of the folds. This latter case involves high glottal airflow, and is characteristic of the utterance-final glottalization that is the topic of this paper. The percept elicited by irregular vocal fold vibration is usually referred to as rough or creaky voice quality.

Glottalization can convey both linguistic and paralinguistic information in normal speech. As an example of its linguistic role, glottalization can serve as an allophone of voiceless stops in American English (particularly syllable-final /t/). Furthermore, it occurs often at intonation phrase onsets as well as at pitch accents, i.e. phrase-level prominences [3,4], if those locations have vowel-initial syllables. Glottalization can also carry paralinguistic information about the speaker's dialect [5], attitude, mood and emotional state [6]. In this paper, we examine another possible paralinguistic function of glottalization: serving as a cue to the identity of the speaker. Glottalization rates vary substantially across speakers; for example, Redi and Shattuck-Hufnagel [7] found that, among their 14 American speakers, one glottalized 88% of the vowel-initial lexical items at intonation boundaries while another one glottalized at only 13%. An earlier study [4] reported glottalization rates for word-initial vowels ranging from 13% to 44% for five professional radio announcers. In Slifka's experiment [2] the four speakers glottalized at the ends of 0%, 51%, 85% and 85% of their vowel-terminal utterances. Slifka notes that the speakers apparently have certain habits in the way they terminate voicing. Although Henton and Bladon [5] do not report quantitative data on individual differences, they note that 10 of the 79 British speakers they examined "spoke in almost continuous creak". Such differences have also been shown for languages other than English. The number of occurrences of glottalization in the same texts varied between 191 and 441 across four Swedish professional speakers [8]. Markó [9] (p. 61) reported that one of her Hungarian speakers frequently glottalized, while the other three seldom did, in recordings of their spontaneous speech.

Because the rate of occurrence of intermittent glottalization seems to be characteristic to at least some speakers, we hypothesize that this voice quality may be one of the acoustic features that listeners utilize to distinguish among familiar talkers, especially for speakers who frequently or seldom glottalize. However, interspeaker differences in an acoustic characteristic do not necessarily imply that listeners use that characteristic in the speaker recognition process. We know that listeners have the ability to recognize a large number of familiar voices from short speech fragments [10], and any or all of the wide range of acoustic parameters that show systematic differences across speakers may serve as a cue for talker recognition. Mean fundamental frequency is believed to be a very robust cue [11,12], but a number of other parameters have been shown to play a role in recognition. In the present study, we conducted a perceptual experiment to determine whether glottalization can contribute to the recognition of familiar speakers. We chose to investigate utterance-final intermittent glottalization in particular because irregularities are very likely to occur at the ends of utterances [5,7] and that it usually has a relatively long time-span in this position, making it acoustically salient.

Our earlier experiment was similar [13], except that the speakers and listeners were all faculty members and graduate students at the same department, and thus were already familiar with each others' voices. This enabled us to investigate the recognition of familiar voices in a straightforward manner, but severely limited the number of potential listeners. To overcome this constraint, the experiment reported here involves perceptual learning of the talkers' voices, allowing us to recruit listeners unfamiliar with our speakers. To determine whether utterance-final voice

quality affects speaker recognition, we created pairs of recordings with regular and irregular endings, and asked whether listeners tend to choose the one with the speaker's usual final voice quality as that speaker's voice. In addition to manipulating utterance-final glottalization, we varied mean F_0 in order to compare the sizes of the effects of these two cue types. This condition also served as a control for the appropriateness of our experimental method for measuring listeners' ability to recognize voices.

2. Method

The experiment consisted of a training phase followed by three test phases. In the training phase, listeners watched a video-recording of each of the four speakers telling a short story. Each of the three test phases began with the presentation of some sentences uttered by the speakers, to remind the listeners of the correspondence between the persons and their voices. Test phases included two sub-tests each: the first sub-test assessed the listener's familiarity with the speakers, and the second used paired comparisons to test the hypothesis that listeners' judgments are influenced by a speaker's habitual voice quality at the ends of utterances. The repetition of the familiarity test in each of the three test phases allowed us to track listeners' progress in learning the voices. The phases of the experiment are summarized in Table 1 and are discussed in more detail in the Procedure subsection below.

2.1. Recordings

Nine American speakers were recorded uttering two tokens each of eight sentences and four individual words and short phrases. The recordings were made directly to a computer at 16 kHz sampling rate using 16 bit quantization, in a sound-treated booth. The ends of all the 216 utterances were labeled as glottalized or non-glottalized by the first author, according to the definition discussed in the Introduction, and the annotation was checked by the second author. These labels were used to calculate the utterance-final glottalization rate for each speaker. As expected, there were some speakers for whom most endings were irregular, others for whom most were regular, and the glottalization rates of the remaining talkers were not as extreme. We selected four speakers for the perceptual experiment: two frequent glottalizers (83% and 93%) and two who seldom glottalized (9% and 20%). Each of these groups included a male and a female.

A video recording was also made of each of the four selected speakers for training purposes. The passage titled 'Comma Gets a Cure' [14] was selected as the script because it contains mainly short sentences and phrases, providing a number of locations that might elicit final glottalization. The video was recorded in a sound-treated booth by a Sony DSC-P150 digital camera with a resolution of 640*480 pixels, at 30 frames per second. The camera also recorded sound, but for presentation purposes this sound was replaced by sound files synchronously recorded by professional equipment at a 16 kHz sampling rate using 16 bit quantization.

2.2. Stimuli for the paired comparisons tests

Each set of stimuli in these tests consisted of an original word uttered by one of the four selected speakers and three manipulated versions of that word. There were 16 such sets, 4 for each speaker, making 64 tokens in total. The original utterance was one of the two recorded versions of each word.

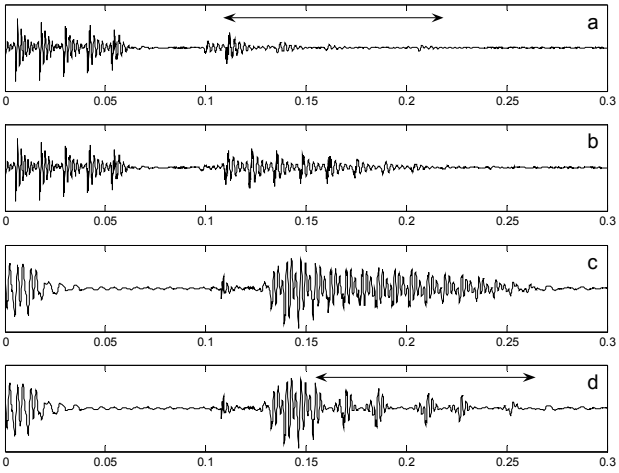

Figure 1. *Examples of unmanipulated recordings (showing only the last 0.3 s) with glottalized (a) and modal (c) endings and their manipulated versions created by concatenation (b) and by cycle attenuation and removal (d). Arrows mark glottalized regions.*

All the words ended with a sonorant. The three manipulations were the following:

1. **Voice quality transformation.** If the end of the original token was produced with regular (modal) phonation, it was amended to sound glottalized (rough) and if it was produced with glottalized (irregular) phonation, it was amended to sound regular. To make the last portion of a modal recording sound glottalized, some of the pitch periods were removed and some others were either attenuated or boosted by a windowing procedure. Fig. 1d shows the result of manipulating the recording in Fig. 1c. To perform the opposite amendment, i.e. to transform a glottalized ending into a modal one, the irregular portion was replaced by a modal ending taken from another utterance (Fig. 1b shows the recording on Fig. 1a transformed to have a modal ending). In two cases there was no such recording available for the speaker, so some pitch periods from the preceding regular region were repeated. The F_0 and amplitude curves of the manipulated endings were shifted up or down to merge smoothly with the preceding regions.
2. **Mean F_0 transformation.** To create these tokens for the higher-pitched female speaker, the F_0 curve of the utterance was shifted down by 30 Hz using Praat [15]. For the lower-pitched female, the F_0 was shifted up by the same amount. A similar transformation was applied to the recordings of the male speakers, but the amount of F_0 shift was ±15 Hz. The F_0 modification was not applied to glottalized regions.
3. **Voice quality and mean F_0 transformation.** To create these tokens, both of the above manipulations were used: first, final voice quality was altered, and then the F_0 contour was shifted.

All stimuli were set to equal rms intensity to minimize loudness differences. The same stimulus set was used in our earlier experiment [13], except that a ±30 Hz F_0 transformation was applied for both the male and the female speakers.

2.3. Listeners

The 11 listeners (6 females, 5 males) were all faculty members and students at universities in the Northeast United States, who reported no speech, language or hearing disorders. They either were native speakers of English or had been living in an English-speaking country for at least three years.

2.4. Procedure

The structure of the experiment is summarized in Table 1.

Table 1. *The stucture of the experiment.*

Phase		Token count
	Training phase	12 minute video
Test phase 1	Reminder	32 sentences
	Familiarity test	48 words
	Paired comparison test	48 pairs
Optional break		
Test phase 2	Reminder	32 sentences
	Familiarity test	48 words
	Paired comparison test	96 pairs
Optional break		
Test phase 3	Reminder	32 sentences
	Familiarity test	48 words
	Paired comparison test	96 pairs

During training, listeners watched a 12 minute video on the computer screen. We used audiovisual training rather than solely auditory exposure because there is evidence that seeing the faces of speakers facilitates the learning of their voices [16]. First, the listener heard the entire recording of the 'Comma' passage by each speaker, and then the last two sentences were replayed. Hearing the same text read by each of the speakers encouraged the listener to pay attention to talker attributes rather than to the content of the message. Each speaker was associated with a short, common first name, and while a speaker was visible, his/her name appeared on the bottom of the screen. The order of the speakers in the video was randomized for each listener.

The training was followed by three test phases, each consisting of three sub-parts: a 'reminder', a familiarity test and a paired comparison test. In the 'reminder', the listener heard eight sentences uttered by each of the four speakers, to refresh their memory. The sentences were blocked by voice, and both the picture and the name of the speaker were displayed on the computer screen.

For the familiarity tests, we used the second original recorded token of each word (not the one that was used to create the three transformed versions for the paired comparisons). Recordings of the same four words uttered by two additional talkers, a male and a female, were also included in the stimulus set to serve as foils. After hearing a recording, listeners were asked to select the speaker from a list of six (names of the 4 known speakers and 'other male', 'other female'). There was no feedback given to the listeners. Each of the 24 recordings (4 words produced by 6 speakers each) was tested twice in randomized order. In our previous

Figure 2: *Screenshot of the paired comparison test. For each judgement, listeners heard a short recording and one of its manipulated versions while the speaker's face and associated (fictional) name were displayed.*

study, listeners also took the familiarity test to determine whether they could reliably recognize the speakers, but they took it only once, at the beginning of the experiment.

For the paired comparison tests, 48 pairs were constructed from the 64 tokens described in the Stimuli subsection, in the following way. One member of the pair was an unmanipulated recording and the other member was a manipulated version of that recording. Thus the pairs differed only in utterance-final voice quality, mean F_0 or both. After hearing a pair, the listener saw the following question on a computer monitor: 'Which one is (or is closer to) X's voice?' where X denoted the name of the speaker. A photo of the speaker was also displayed in order to avoid difficulties in name recall. Listeners gave their answers by clicking on a 6-point scale displayed on the screen, where button 1 was labeled 'Certainly the first' and button 6 as 'Certainly the second'. See Figure 2 for a screenshot of the user interface of the test. During the first test phase, each pair was tested once and half of the pairs (selected randomly) were presented in reversed order, i.e. with the unmanipulated recording as the second token of the pair. In the second and third test phases, the pairs were tested four times (yielding 96 trials in each phase): the unmanipulated recording occurred twice as the first token of the pair and twice as the second. Presentation order was re-randomized for each listener and for each test phase. In our earlier experiment, there were only two paired comparison tests with 96 stimulus pairs each. The additional, shorter test included here in phase 1 can be considered practice, if the results of the familiarity assessment indicates that, at this point in the experiment, listeners have not yet learnt the voices.

The three test phases were separated by optional breaks. Listeners were tested individually in a quiet office, using a PC and Bose TriPort II headphones. The test was administered using a graphical program written in Matlab 7.1. Listeners produced their responses by clicking the appropriate button on the screen using the mouse. There was no time limit for giving a response, but listeners were instructed to respond quickly. An experimental session lasted about 45 minutes.

3. Results[1]

3.1. Familiarity tests

We compared the results for the three familiarity tests with those of our earlier experiment [13]. As noted above, the familiarity test was the same as the one used here (except that it was taken only once rather than three times), so it was possible to compare the data from the two experiments.

A one-way analysis of variance (ANOVA) on the correctness of the responses showed a significant effect of test phase ($F = 8.318$; $p < 0.0005$). Data from the previous study was considered as a fourth phase. According to Tukey's post-hoc tests, the recognition rate increased from the first to the second phase ($p = 0.006$; Fig. 3). The results of the second and third phases were not significantly different from each other ($p = 0.836$, not significant) or from the recognition rate achieved by the familiar listeners in the other experiment ($p = 0.625$ and $p = 0.981$, n.s.). This suggests that listeners benefited from the second 'reminder' or that perhaps the time elapsed between the initial training and the second test helped to consolidate the new knowledge. In the second and third phase however, listeners achieved a similar recognition rate to that shown by the listeners who were already familiar with the test voices. This suggests that by the second test phase they had acquired the information needed to recognize the four voices. Consequently, we considered the first phase to be practice, and excluded the paired comparison responses of this phase from further data analysis.

The effectiveness of the training was determined for each listener by comparing recognition rates to chance performance. Although there were six possible responses, chance level was considered to be 33% since gender recognition was perfect. One-sample t-tests showed that recognition rates were significantly higher than chance for nine listeners in the first test and for all eleven listeners in both the second and the third test ($t \geq 2.331$; $p \leq 0.024$). This means that all of the listeners became at least somewhat familiar with the speakers by the second phase. The recognition rates of nine out of the eleven listeners improved during the experiment and for the other two listeners it remained roughly the same. The recognition rates varied much less across listeners in this experiment than in the earlier one: the range of the individual results here was 53-75% in contrast to the 46-83% in the other study. One reason behind this difference may be that listeners' familiarity was more controlled here: they all had the same amount and means of exposure to the speakers' voices. As in the previous experiment, recognition rates for the two familiar female speakers, especially for the female frequent glottalizer, were lower than for the two familiar males, suggesting that the voice of the female "glottalizer" is harder to identify.

3.2. Paired comparisons

One listener (Listener 3) pressed the same button 95% of the time throughout the two paired comparison tests, and was also one of the two subjects whose recognition rate did not improve across the three familiarity tests. Accordingly, his responses were withdrawn from the data set, and we conducted all further analyses for the remaining ten listeners only.

[1] All the analysis was carried out at 5% level.

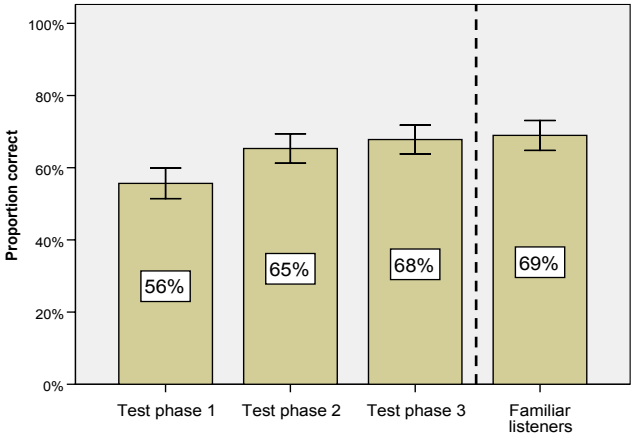

Figure 3: *Proportion of correct responses in the three familiarity tests, compared with the results of familiar listeners in an earlier experiment. Error bars represent 95% confidence intervals.*

When listeners chose the original rather than the manipulated recording, we considered it a correct response; otherwise it was incorrect. Thus, correctness was a measure of how effectively listeners could use the different cue types in recognizing the speaker, or, to put it another way, how effectively changing a cue interfered with the listeners' ability to identify the voice. We also extracted from the responses how confident listeners were in their choice (low: 3, 4; mid: 2, 5; high: 1, 6). Confidence ratings were significantly higher for correct responses than for incorrect responses ($t = -9.640$; $p < 0.0005$).

Fig. 4 shows the proportion of correct responses for the three experimental conditions. When utterance-final voice quality was manipulated, 60% of the responses were correct. That is, tokens with the original voice quality were preferred over tokens with changed voice quality 60% of the time. A one-sample t-test showed that this is significantly higher than the 50% chance level ($t = 5.328$; $p < 0.0005$), indicating that changing voice quality made the speaker less identifiable.

As expected from the fact that mean F_0 has been shown to be a useful cue to the speaker [11], the correct response rate is higher (72%) for the case when the F_0 contour was shifted up or down for the transformed member of the pair. Changing the average F_0 had a stronger effect than changing final glottalization, but altering the voice quality still affected listeners' decisions significantly.

An ANOVA was conducted on correctness with condition and speaker as fixed factors and listener as random factor. The main effect of condition was significant ($F = 19.333$; $p < 0.0005$) and Tukey's post-hoc tests showed a significant difference between the correct response rates of condition *glott* and F_0 ($p < 0.0005$) but not between F_0 and *glott*+F_0 (where both voice quality and mean F_0 were changed; $p = 0.217$, n.s.).

The significant interaction between speaker and listener ($F = 6.546$; $p < 0.0005$) indicated that listeners differed in how hard they found the task for different speakers. This is consistent with the results of the familiarity tests, which showed large variation among both listeners and speakers in their recognition rates.

The condition-by-listener interaction was at the boundary of significance ($F = 2.106$; $p = 0.050$, n.s.). If this interaction becomes significant as we recruit more subjects (as was the

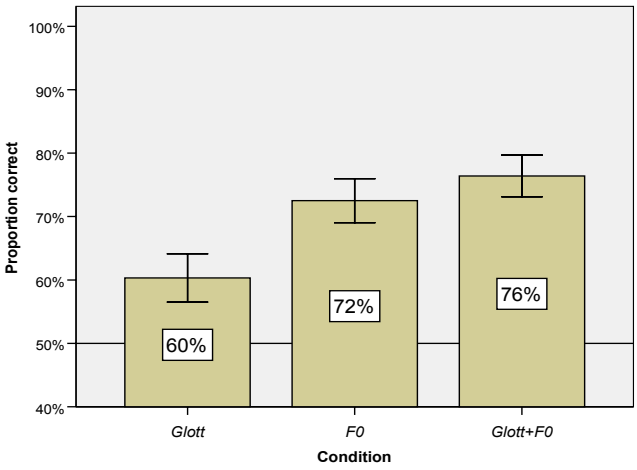

Figure 4: *Proportions of correct responses in the paired comparison tests. Results pooled by condition, i.e. one member of the stimulus pair was manipulated by altering utterance-final voice quality (glott), mean F_0 (F0) or both (glott+F0). The horizontal line corresponds to the 50% chance level and error bars represent 95% confidence intervals.*

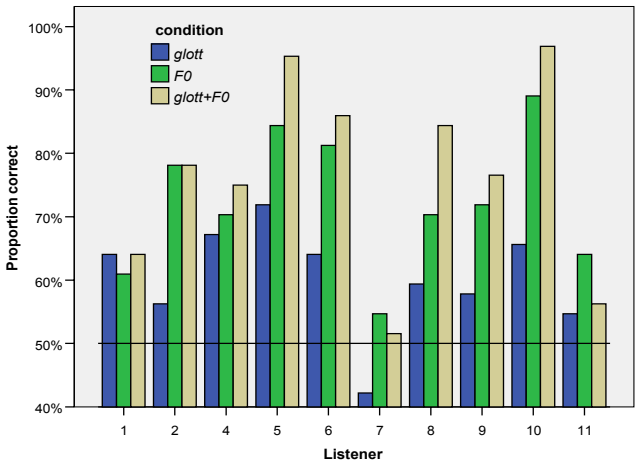

Figure 5: *Proportions of correct responses in the paired comparison tests by listener and by condition. The horizontal line corresponds to the 50% chance level.*

case in our previous experiment), it will support the idea that different listeners utilize different cues in recognizing a voice. For the *glott* condition, the rate of correct responses ranged from 42% to 72% (Fig. 5). For example, Listener 2 with a near-chance performance for the *glott* condition showed an above-average correct rate for the F_0 condition. On the other hand, some listeners with a high correct percentage for the *glott* condition (Listener 1 and Listener 4) achieved a roughly similar score for the F_0 condition also.

4. Summary

According to previous studies, there are some speakers who produce intermittent episodes of glottalization regularly and some who seldom do so. Thus, these regions of irregular pitch periods in certain locations may be one of the acoustic parameters employed in recognizing a familiar speaker. A perceptual experiment is being conducted to test whether the presence or absence of glottalization at the ends of utterances affects speaker recognition. This experiment involves training the listeners to recognize the target voices and it is a follow-up to an earlier experiment [13] in which listeners were already familiar with the speakers' voices.

Our initial results with a limited number of listeners receiving this training are encouraging and are in line with our earlier results for listeners already familiar with the test voices. They show that listeners encode this information about the talker (i.e. whether the speaker is a frequent vs. a rare glottalizer) in memory and can access it and make use of it in this task: when they heard pairs of speech samples they tended to choose the one with the speaker's usual utterance-final voice quality as the one that was closer to the speaker's voice. Similar results were obtained in our earlier study (where listeners prefered the speaker's habitual voice quality in 63% of the time compared to 60% in this experiment). In our earlier study the rate of correct responses was 85% for the F_0 condition and 90% for *glott*+F_0. The lower percentages for these two conditions in the experiment reported here (72% and 76%) can be explained by the smaller F_0 shift applied to male speech (±15 Hz instead of ±30 Hz). In the earlier experiment, the roughly same difference of 5 percentage points was measured between the F_0 and *glott*+F_0 conditions but in that study the difference was significant, indicating that having both the glottalization cue and the mean F_0 cue made it easier for familiar listeners to judge which of the two speech samples was produced by the speaker.

In contrast to the several-day training methods reported in the literature (e.g. in [17]), our simple familiarization procedure was applied in a single session. Nevertheless, it was effective in this context: even though they watched and listened only passively for less than half hour and received no feedback, our listeners achieved recognition rates that are similar to those of listeners who were already familiar with the speakers. In the familiarity tests, the variation across speakers and listeners was large. This may mean that speaker recognition performance depends on both the speaker and the listener, as was suggested by Van Lancker et al. [18]. Taken together with previous results in the literature, these observations support the hypothesis that, in recognizing familiar voices, listeners make use of a speaker's characteristic pattern of intermittent change in voice quality.

5. Acknowledgements

The authors are grateful to Géza Németh for his insights concerning the voice quality transformation and to the Embassy of the United States in Budapest, Hungary for help in recruiting listeners. The first author was partially funded by the Hungarian National Office for Research and Technology grant NKFP 2/034/2004, and the second author by NIH grants RO1-DC002978 and RO1-DC0075.

6. References

[1] Laver, J. The phonetic description of voice quality, Cambridge University Press, Cambridge, 1980.

[2] Slifka, J. "Some physiological correlates to regular and irregular phonation at the end of an utterance", J. Voice 20:171-186, 2006.

[3] Pierrehumbert, J. and Talkin, D. "Lenition of /h/ and glottal stop", Papers in Laboratory Phonology II: Gesture, Segment, Prosody, D. Docherty and D.R. Ladd,

Eds. Cambridge University Press, Cambridge, 1992, pp. 90-117.

[4] Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M. “Glottalization of word-initial vowels as a function of prosodic structure”, J. Phonetics 24:423-444, 1996.

[5] Henton, C.G. and Bladon, A. “Creak as a sociophonetic marker”, in Language, speech and mind: Studies in honour of Victoria A. Fromkin, L.M. Hyman and C.N. Li, Eds. Routledge, London, 1987, pp. 3-29.

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

[7] Redi, L. and Shattuck-Hufnagel, S. “Variation in the realization of glottalization in normal speakers”, J. Phonetics 29:407-429, 2001.

[8] Hedelin, P. and Huber, D. “Pitch period determination of aperiodic speech signals”, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Albuquerque, pp. 361-364, 1990.

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

[10] Van Lancker, D., Kreiman, J., and Emmorey, K. “Familiar voice recognition: patterns and parameters; Part I”, J. Phonetics 13:19-38, 1985.

[11] Abberton, E. and Fourcin, A.J. “Intonation and speaker identification”, Language and Speech 21:305-318, 1978.

[12] Kreiman, J. and Papcun, G. “Comparing discrimination and recognition of unfamiliar voices”, Speech Communication 10:265-275, 1991.

[13] Bőhm, T. and Shattuck-Hufnagel, S. “Utterance-final glottalization as a cue for familiar speaker recognition”, Proc. Interspeech, Antwerp, 2007, in press.

[14] Honorof, D.N. (Ed.), McCullough, J., and Somerville, B. “Comma Gets A Cure” [diagnostic passage], 2000, retrieved June 10, 2006, from http://web.ku.edu/idea/readings/comma.htm

[15] Boersma, P. and Weenink, D. “Praat: doing phonetics by computer (Version 4.4.10)” [Computer program], retrieved February 22, 2006, from http://www.praat.org/

[16] Sheffert, S.M. and Olson, E. “Audiovisual speech facilitates voice learning”, Perception & Psychophysics 66:352-362, 2004.

[17] Nygaard, L.C. and Pisoni, D.B. “Talker-specific learning in speech perception”, Perception & Psychophysics 60:355-376, 1998.

[18] Van Lancker, D., Kreiman, J., and Wickens, T.D. “Familiar voice recognition: patterns and parameters; Part II”, J. Phonetics 13:39-52, 1985.