# Utterance-Final Glottalization as a Cue for Familiar Speaker Recognition

*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

10.21437/Interspeech.2007-698

## Abstract

Several studies have reported systematic differences across speakers in the rate and type of intermittent irregular vocal fold vibration (glottalization). Still, it remains an open question whether human listeners use this speaker-specific information as a cue for recognizing familiar voices. A perceptual experiment was conducted to investigate this issue, concentrating on irregularity in utterance-final position. A novel method was employed to manipulate the final voice quality (in our case, modal or glottalized). Listeners, who were familiar with the voices of the speakers, were presented pairs of speech samples: one with the original and another with manipulated final voice quality. When listeners were asked to select the member of the pair that was closer to the talker's voice, they chose the unmanipulated token in 63% of the trials. This result suggests that irregular pitch periods in utterance-final regions play a role in the recognition of individual speaker voices.

**Index Terms**: speaker recognition, memory for voices, glottalization, creak, voice quality

## 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. Fig. 1a shows an example of glottalization occuring at the end of an utterance. The perceivability criterion serves to exclude minor deviations from periodicity that are inherent to human phonation. The irregularity may occur in the time spacing of the glottal pulses, in their amplitude or in both of these parameters. It is often accompanied by full damping of the pulses and low $F_0$. The present study focuses on intermittent glottalization, i.e. glottalized regions in otherwise modal speech, in contrast to persistent glottalization in which a speaker's phonation is consistently irregular. Note that classifying a speech fragment as glottalized or not glottalized in this sense is often somewhat challenging, especially when the ambiguous fragment is short. In this paper, however, we investigate utterance-final glottalization, which usually spans a longer time period than glottalization in other positions, making it more clearly distinct from non-glottalized endings.

Although this definition covers a variety of acoustic manifestations (e.g. all the four categories of vocal aperiodicity discussed in [1]), it does not assume anything about the underlying production or perception mechanisms (except perceivability). 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 [2] (p. 122-126). Recently, an experiment involving simultaneous acoustic and physiological measurements by Slifka [3] showed that irregular vocal fold vibration can also be produced by the abduction of the folds. This latter case involves high glottal airflow and it is characteristic of 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. In Catford's informal description, the auditory effect is often "like a stick being run along a railing" (cited in [2], p. 122).

Persistent or unwanted glottalization may be a symptom of vocal disorders, but glottalization often occurs intermittently in normal speech as well [4], where it can play a communicative role. For example, in American English glottalization (or glottal stop) can serve as an allophone of voiceless stops (particularly syllable-final /t/), and it often occurs at the onset of vowel-initial words that begin a new intonation phrase or carry a pitch accent (i.e. a phrase-level prominence) [5].

Several studies have shown substantial individual differences in the rate of occurrence of this intermittent voice quality. Redi and Shattuck-Hufnagel [6] found that, among their 14 American speakers, one glottalized 88% of the regions examined while another glottalized only 13%. An earlier study [7] reported glottalization rates for word-initial vowels ranging from 13% to 44% for five professional radio announcers. In Slifka's experiment [8] (p. 100-103) the four speakers glottalized at the ends of 5%, 37%, 93%, and 95% of their utterances. Slifka notes that speakers appear to have certain habits in the ways they terminate voicing. Although Henton and Bladon [9] do not report quantitative data on individual differences, they note that 10 of the 79 British speakers they examined showed glottalization in almost all the syllables, while some others showed it in only a few. Between-speaker differences have also been shown for languages other than English. For example, tokens of glottalization varied between 191 and 441 across four Swedish professional speakers [1], and Markó [10] (p. 61) reported that one of her Hungarian speakers frequently glottalized, while the other three seldom did, in recordings of their spontaneous speech.

Because intermittent glottalization seems to be characteristic for at least some speakers, we hypothesize that this voice quality may be one of the acoustic features that listeners use to distinguish among familiar talkers, especially for speakers who frequently or seldom glottalize. However, interspeaker differences in an acoustic feature do not necessarily imply that they are used by human listeners to recognize speakers. We know that listeners have the ability to recognize a large number of familiar voices from short speech fragments [11], and any of the wide range of acoustic parameters that show systematic differences across speakers may serve as a cue for talker recognition. Pitch and pitch range are believed to be the most robust ones [12], but a number of other parameters have been shown to play a role [11]. In the present study, we tested whether glottalization is one of these cues for familiar speakers. We conducted a perceptual experiment to determine whether listeners retain such information in their memory representation of a talker. The fact that intermittent glottalization is likely to occur at the ends of utterances [6,9], and that in this location it usually has a relatively long time-span (making it acoustically salient), led us to focus on this position. We created pairs of recordings with regular and irregular endings, and asked whether listeners tend to choose the one with the speaker's typical final voice quality as that speaker's voice. To create these pairs we manipulated utterance-final glottalization, and we also varied the mean $F_0$ which is a very robust cue for speaker identification. In this way we could compare the effectiveness of these two cue types, and also control for the appropriateness of our experimental method.

# 2. Method

## 2.1. Stimuli

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 a 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, for some speakers most endings were irregular, for some other speakers 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%). Both groups consisted of a male and a female.

Each set of stimuli consisted of an original word or phrase uttered by one of the four speakers and three manipulated versions of that word. All the words ended with a sonorant. There were 16 such sets, 4 for each speaker, making 64 tokens in total. The original utterance with the speakers' most typical final voice quality was selected from the two recorded versions of each word. The three manipulations were:

1. **Voice quality transformation.** If the end of the original token was produced with regular phonation, it was amended to sound glottalized (rough) and vice versa. In order to make the last portion of a modal (non-irregular) recording sound glottalized, some of the pitch periods were zeroed out 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 connect smoothly with the preceding regions. A formal evaluation of the two voice quality transformation methods showed that converted speech approached natural glottalized and modal utterance-endings (respectively) in terms of perceived roughness with no significant degradation in naturalness. These results are described in [13].

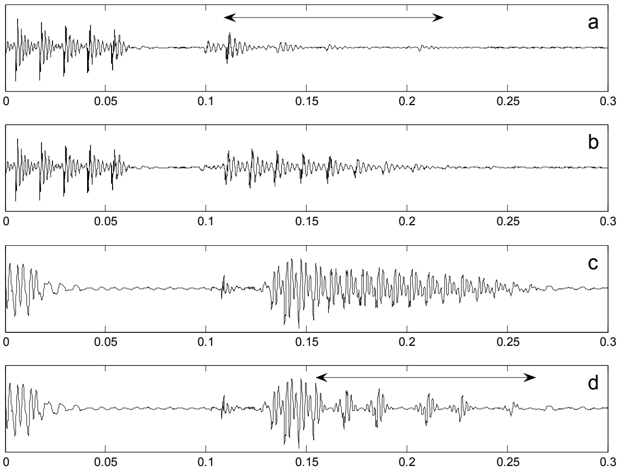

Figure 1. *Examples of unmanipulated recordings (showing only the last 0.3 s) with glottalized (a) and modal (c) endings and their manipulated versions created by concatenation (b) and cycle removal (d). Arrows mark glottalized regions.*

2. **Mean $F_0$ transformation.** For the higher-pitched male and female speaker, the $F_0$ curve of the utterance was shifted down by 30 Hz using Praat. For the lower-pitched male and female, the $F_0$ was shifted up by the same amount. The $F_0$ modification was not applied to glottalized regions.
3. **Voice quality and mean $F_0$ transformation.** In this case, both of the above manipulations were used: first, final voice quality was altered, and then the pitch contour was shifted.

The stimuli were set to equal intensity based on an RMS-measure, to minimize loudness differences between the members of the pairs.

## 2.2. Listeners

The 10 listeners (4 females, 6 males) were all faculty members or graduate students at the department that the four speakers were affiliated with. By self-report they were familiar with all the speakers' voices. The listeners were either native speakers of English or had been living in an English-speaking country for at least three years.

## 2.3. Procedure

The experiment consisted of two tasks: the first assessed the listener's familiarity with the speakers, and the second used paired comparisons to determine whether listeners remember a speaker's habitual voice quality at the ends of utterances

For the familiarity test, we used the second original recorded token of each word (not the one used to create the three transformed versions). Recordings of the same four words uttered by a male and a female talker unknown to the listeners were included in the stimulus set as foils. After hearing a token, listeners were asked to select the speaker from a list of six (names of the 4 known speakers and 'other male', 'other female'). Each of the 24 recordings (4 words produced by 6 speakers each) was presented twice in randomized order.

For the second test, 48 pairs were constructed from the 64 tokens described in subsection 2.1. in the following way. One member of the pair was an unmanipulated recording and the other one 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. 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’. Each pair was tested four times (yielding 192 trials): 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.

Listeners were tested individually in a quiet office, using a PC and Bose TriPort II headphones. The test was administered via a graphical program written in Matlab 7.1. Responses were given by clicking on the appropriate button on the screen using the mouse.

# 3. Results

## 3.1. Familiarity test

In the familiarity test, listeners recognized the speaker correctly on 69% of the trials. Although there were six possible responses, chance level was considered to be 33% since gender recognition was perfect. For the discussion below, we adopt the significance criterion of $p < 0.05$. One-sample t-tests showed that recognition rates were significantly higher than chance for nine listeners ($t \geq 2.331$; $p \leq 0.024$). The recognition rate of the remaining one listener (46%) was still well above chance and fell just short of significance ($p = 0.084$). Although the recognition rates of the listeners vary, all of them can be considered to be familiar with the speakers. The 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 for these listeners.

## 3.2. Paired comparisons

To analyse these results, for each response we determined whether listeners chose the original rather than the manipulated recording. When they did so, we considered it a correct response. When they selected the manipulated token, it was considered incorrect. Thus, correctness was a measure of how effectively listeners used the different cue types in recognizing the speaker. 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 signficantly higher for correct responses than for incorrect responses ($t = 22.879$, $p < 0.0005$).

Fig. 2 shows the proportion of correct responses for the three experimental conditions. When utterance-final voice quality was manipulated, 63% of the responses were correct. That is, tokens with the original voice quality were preferred over tokens with changed voice quality 63% of the time. A one-sample t-test showed that this is significantly higher than the 50% chance level ($t = 6{,}789$; $p < 0.0005$), indicating that changing voice quality made the speaker less identifiable.

As expected (since mean $F_0$ has been shown to be a robust cue to the speaker [12]), the correct response rate is high (85%) for the cases where the $F_0$ contour was shifted up or down for the transformed member of the pair. That is, changing the average $F_0$ had a stronger effect than changing final glottalization. However, altering the voice quality did affect listeners’ decisions significantly, even though to a lesser extent. The significant increase in the rate of correct responses for the *Glot+F0* condition (where both voice quality and mean $F_0$ were changed) compared to the *F0* condition ($t = -2.99$; $p = 0.003$) indicates that having both the *Glot* cue and the *F0* cue makes it easier for listeners to tell which of two speech samples was produced by the speaker.

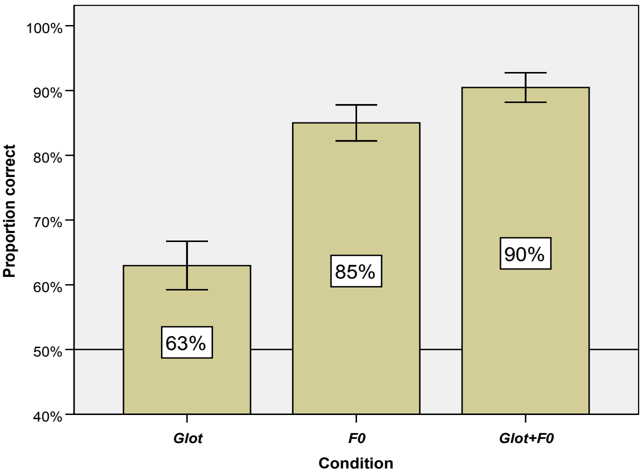

Figure 2: *Proportions of correct responses by condition, i.e. one pair member was manipulated by altering utterance-final voice quality (Glot), mean F0 (F0) or both (Glot+F0). The horizontal line corresponds to the 50% chance level and error bars represent 95% confidence intervals.*

An analysis of variance (ANOVA) was conducted for correct responses, with condition and speaker as fixed factors and listener as random factor. The significant interaction between condition and speaker ($F = 5.509$; $p < 0.0005$) showed that the two cues may carry different weights for recognizing some speakers than for others (Fig. 3). For example, when the presence or absence of final glottalization was varied for the “nonglottalizer” male speaker, 75% of the responses were correct while this rate was 57-60% for the other three speakers. For the female “glottalizer” (who was harder to recognize, according to the familiarity test), the proportion of correct responses for the *Glot* and *F0* conditions differed much less than for the other speakers (59% and 69%, respectively).

The significant interaction between condition and listener ($F = 2.508$; $p = 0.005$) supports the idea that different listeners utilize different cues in recognizing a voice. Some listeners with a near-chance performance for the *Glot* condition showed an above-average correct rate for the *F0* condition (Fig. 4). On the other hand, listeners with the highest correct percentage for the *Glot* condition achieved a roughly similar score for the *F0* condition also. For the *Glot* condition, the rate of correct responses ranged from 50% to 83%.

The speaker-by-listener interaction and the main effect of listener were not significant ($F = 1.642$ and $F = 1.038$), so the results cannot be explained by differences among listeners’ degree of familiarity with the speakers’ voices. We do not report the main effects of condition, speaker and listener because we showed significant interactions among them.

# 4. Summary

According to previous studies, there are some speakers who produce intermittent glottalization regularly and some who produce it seldom. Thus, intermittent episodes of irregular pitch periods in certain locations may be one of the acoustic parameters employed in recognizing a familiar speaker. Our results show that listeners encode information about the talker's likelihood of glottalizing utterance-finally (frequent vs. rare) in memory: from pairs of speech samples they tended to choose the member with the speaker's typical utterance-final voice quality as the one that was closer to the speaker's voice. The large variation across speakers and listeners suggests that the weight of this cue may be different by speaker and by listener. Van Lancker et al. [14] also found that the set of cues critical to voice recognition is a function of both the speaker and the listener, and Kreiman et al.'s results [15] on the variation in voice quality ratings further support listener-specificity.

Requiring that potential listeners be already familiar with the voices in the experiment severely restricts their number. A subsequent experiment involving within-experiment perceptual learning of talkers' voices has enabled us to recruit from a wider population. Preliminary results using this method are very close to those reported here, and also to a similar previous experiment using formant-synthesized stimuli [16]. Together with previous results in the literature, these observations support the hypothesis that listeners make use of a speaker's characteristic pattern of intermittent change in voice quality in recognizing familiar voices.

## 5. Acknowledgements

The authors are grateful to Kenneth N. Stevens for guidance in experiment design, Mária Gósy for helpful insights, and István Maricza for assistance in data analysis. The first author was supported by a Fulbright Scholarship in the initial stage of this work and 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.

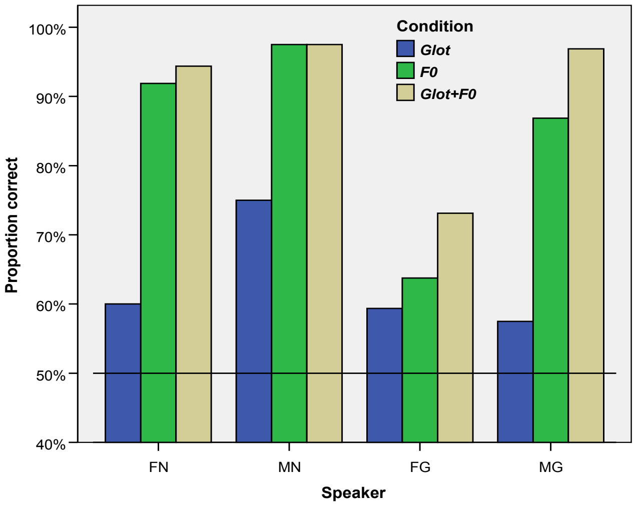

Figure 3: *Proportions of correct responses by speaker and by condition. Speakers are identified by gender (M/F) and by 'glottalizer'/'non-glottalizer' (G/N). The horizontal line corresponds to the 50% chance level.*

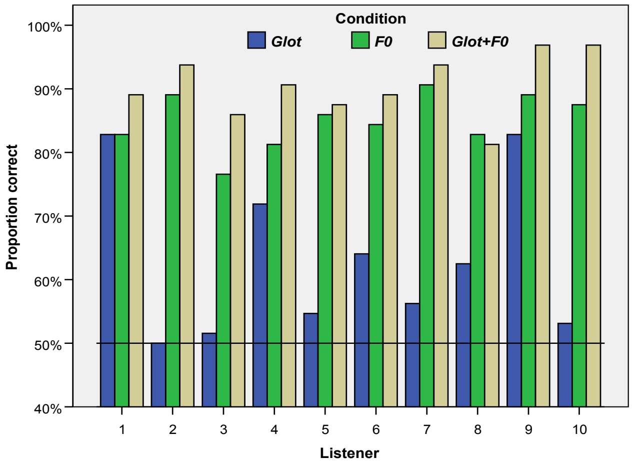

Figure 4: *Proportions of correct responses by listener and by condition. The horizontal line corresponds to the 50% chance level.*

## 6. References

[1] 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.
[2] Laver, J. The phonetic description of voice quality, Cambridge University Press, Cambridge, 1980.
[3] Slifka, J. "Some physiological correlates to regular and irregular phonation at the end of an utterance", J. Voice 20:171-186, 2006.
[4] Hollien, H. and Wendahl, R.W. "Perceptual study of vocal fry", J. Acoust. Soc. Amer. 43:506-509, 1967.
[5] 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.
[6] Redi, L. and Shattuck-Hufnagel, S. "Variation in the realization of glottalization in normal speakers", J. Phonetics 29:407-429, 2001.
[7] 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.
[8] Slifka, J. Respiratory constraints on speech production at prosodic boundaries, Ph.D. dissertation, MIT, Cambridge, 2000.
[9] 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.
[10] Markó, A. A spontán beszéd néhány szupraszegmentális jellegzetessége, Ph.D. dissertation, ELTE, Budapest, 2005.
[11] Van Lancker, D., Kreiman, J., and Emmorey, K. "Familiar voice recognition: patterns and parameters; Part I", J. Phonetics 13:19-38, 1985.
[12] Abberton, E. and Fourcin, A.J. "Intonation and speaker identification", Language and Speech 21:305-318, 1978.
[13] Bőhm, T., Shattuck-Hufnagel, S., and Németh, G. "A simple method to convert modal into irregular voice", in preparation.
[14] Van Lancker, D., Kreiman, J., and Wickens, T.D. "Familiar voice recognition: patterns and parameters; Part II", J. Phonetics 13:39-52, 1985.
[15] Kreiman, J., Gerratt, B.R., Precoda, K., and Berke, G.S. "Individual differences in voice quality perception", J. Speech and Hearing Res. 35:512-520, 1992.
[16] Bőhm, T. "Is utterance-final glottalization a cue for speaker recognition by humans?", 151st Meeting of the Acoustical Society of America, Providence, 2006.