1pSC24. **Is utterance-final glottalization a cue for speaker recognition by humans?**

Tamás Bőhm

Department of Telecommunications and Media Informatics, BME, Budapest, Hungary
and Research Laboratory of Electronics, MIT, Cambridge, MA, bohm@tmit.bme.hu

Background

Definition

Glottalization is defined here as irregular vocal fold vibration.

- Irregular in spacing or amplitude of the glottal pulses or both.
- It does not refer to any particular production mechanism.
- It often shows a distinctive pulse decay pattern.

Previous studies

Nonmodal phonation, including glottalization, can be perceived and discriminated by human listeners at least in some circumstances (Huber, 1992).

Several studies reported that the rate of occurence and acoustic characteristics of glottalization varies substantially across speakers (Huber, 1990; Dilley et al., 1996; Redi and Shattuck-Hufnagel, 2001).

The speakers in Slifka's (2001) investigation produced glottalization at the ends of utterances with a consistently different rate of occurence. Speakers appeared to have certain habits for terminating voicing: some produced mainly regular endings while others mainly irregular ones.

Research question

Do listeners exploit these interspeaker differences? Do they use utterance-final voice quality changes to recognize individual voices?

The answer can guide the development of speech technologies such as

- voice converters
- personalized text-to-speech sytems (that aim to quickly learn new voices).

Hypotheses

1. Utterance-final glottalization is perceiveable by listeners.
   - It has been shown for synthetic sustained vowels that may not generalize for running speech and for natural speech though not controlled for other variation (Huber, 1992).
   - We tried to get more direct evidence by varying only this aspect of the utterances and using meaningful words.
2. Utterance-final glottalization helps human listeners to recognize familiar voices (for some speakers and some listeners).

Method

A listening experiment was conducted to test the hypotheses. The outline of the method:

1. Select some speakers who reliably glottalize and some who reliably do not and record a set of utterances from both sets of speakers.
2. Create two copy-synthesis versions of selected utterances: one with modal and one with glottalized voice quality in their final regions.
   ➔ For each speaker, one of the conditions reflects the speaker's usual utterance-final voice quality and the other does not.
3. Determine which one of each pair of stimuli do listeners judge to be the speaker's voice.
   - Are the answers random or lean towards one of the conditions?
   - If listeners tend to choose the condition that is characteristic of the speaker then it supports the hypotheses.

As a basis of comparison for utterance-final glottalization, another factor (pitch contour) was varied: in half of the cases the time-wrapped pitch contour of another speaker was used because this parameter is widely considered to be a strong perceptual cue on speaker identity.

Stimuli

Original utterances

4 speakers were selected from a pool of 9, based on glottalization rates at final regions:

- 2 frequent 'glottalizers' (86% and 93%)
- 2 infrequent 'glottalizers' (9% and 20%)
- Both groups consisted of 1 female and 1 male.

4 short utterances by each speaker (16 recordings in total) were selected for copy-synthesis using the Klatt synthesizer.

Experimental manipulation

Two factors were varied for the copy-synthetic utterances:

G: presence/absence of final irregularity is original or 'inverted'

- Glottalization was synthesized (Figure 1) by abrupt changes in fundamental frequency and amplitude of voicing. Other parameters frequently manipulated: open quotient, spectral tilt, flutter, degree of diplophonia, and delta B1 (increase of B1 during open portion of the glottal period).
- The criterion was to get a close match in both perceptual and acoustic terms (informal listening and spectrogram comparison).

P: pitch contour (original / the one belonging to the other speaker of the same gender)

All the 16 utterances were synthesized with all the 4 possible combinations of these two factors (Figure 2), resulting in 64 tokens.

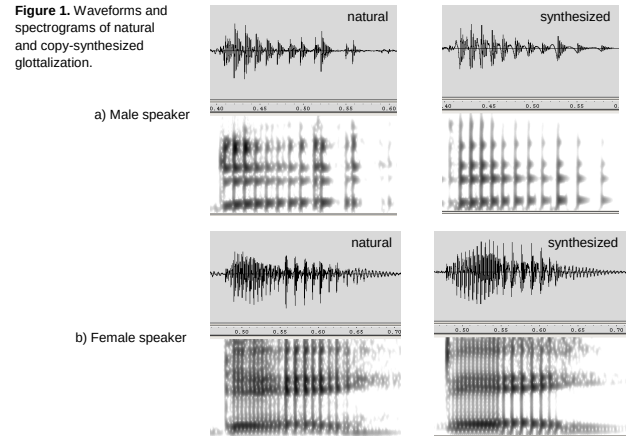


Figure 1. Waveforms and spectrograms of natural and copy-synthesized glottalization.

Glottalization and pitch as in original

Glottalization inverted (regular ending for 'glottalizers' and irregular ending for 'nonglottalizers')

Pitch contour copied from another speaker

Both glottalization and pitch contour modified

Figure 2. The four conditions in which each recording was synthesized.

Testing procedure

Subjects

- 6 subjects (4 females, 2 males), graduate students, all familiar with the 4 selected voices
- All unaware of the hypotheses and the nature of the stimuli.

First part of the experiment: it checked…

…if the listener is familiar with the voices
…if the copy synthetic utterances sound like the target speakers
…if the synthesis of glottalization sounds glottalized

Second part of the experiment: tested the hypotheses -- if utterance-final glottalization is perceiveable and if it helps listeners to identify voices -- with pair comparison (AB test). Both members of a pair were based on the same utterance by the same speaker but they represented a different condition (e.g., one of them was modal at the end while the other was glottalized; they were otherwise identical).

After hearing a pair of stimuli, listeners were asked "Which one is (or closer to) X's voice?" where X is the name of the speaker. Answers were given on a 6-point scale with the extremes labeled. '1' meant 'certainly the first' utterance of the pair and '6' corresponded to 'certainly the second'.

Results

Analysis: Preliminary analysis of the data was carried out. In this analysis, answers were considered binary: responses within the range '1'-'3' were treated as identical, as were responses within the '4'-'6' range.

Effect of glottalization: For the pairs where only the glottalization pattern was varied (i.e. one member is glottalized, the other is modal), listeners chose the one that represented the speaker's usual final voice quality 63% of the times (Figure 3), that is significantly ($\alpha < 0.05$) different than the 50% chance level. This result supports both of the hypotheses.

Effect of pitch: Figure 4 shows the answer means for all the pairs. Pitch contour is definitely a stronger cue of speaker identity than final glottalization. Still, the means for all the types of pairs are compatible with the hypotheses.

Figure 3. Mean of listeners' answers (right) for the type of pairs depicted on the left.

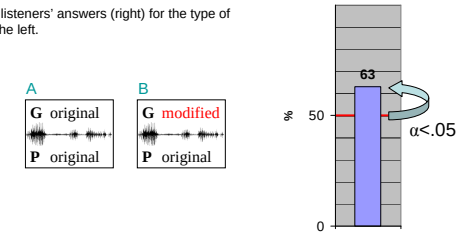
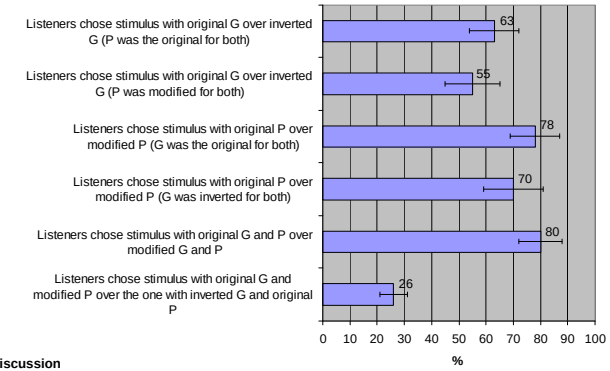


Figure 4. Means of listeners' answers for all the six types of pairs.



Discussion

The results support both of the hypotheses, namely:

- Listeners were able to distinguish glottalized speech from non-glottalized.
- Utterance-final glottalization seems to play a role in the perception of speaker identity, but a finer analysis of the results is necessary.

Multiple cues

Utterance-final glottalization appears to be one of the several cues used in the human speaker recognition process, at least for some speakers and some listeners.

Future experiments

Further experiments should use more realistic stimuli. Instead of copy-synthetic utterances, manipulated natural stimuli may be considered.

Further research should investigate glottalization at other positions (such as utterance initial) and the role of other voice qualities. Also, synthesis methods of glottalization need further study.

Acknowledgements

The author is grateful to Kenneth N. Stevens and Stefanie Shattuck-Hufnagel for their helpful insights and guidance in designing the experiment. The majority of this work was carried out when the author was a visiting student at the Research Laboratory of Electronics, MIT (Cambridge, MA) and while he was supported by a Fulbright Fellowship.

References

L. Dilley, S. Shattuck-Hufnagel, M. Ostendorf (1996): Glottalization of word-initial vowels as a function of prosodic structure, Journal of Phonetics, vol. 24, pp. 423-444.

P. Hedelin, D. Huber (1990): Pitch period determination of aperiodic speech signals, Proceedings of ICASSP 1990, pp. 361-364.

D. Huber (1992): Perception of aperiodic speech signals, Proceedings of ICSLP 1992, pp. 503-506.

L. Redi, S. Shattuck-Hufnagel (2001): Variation in the realization of glottalization in normal speakers, Journal of Phonetics, vol. 29, pp. 407-429.

J. Slifka (2000): Respiratory Constraints on Speech Production at Prosodic Boundaries, PhD thesis, MIT.