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, bohm@tmit.bme.hu
[2]Research Laboratory of Electronics, MIT, Cambridge, MA USA, stef@speech.mit.edu

MŰEGYETEM 1782

Massachusetts Institute of Technology

Familiar speaker recognition

- How does a listener recognize a familiar speaker?
- One possibility is a featural model:

speech → Feature extraction → features → Recognition → identity

- What are the features?
  - A speaker's voice carries many characteristic acoustic parameters (e.g. mean F_0)
  - Listeners may not use all of these speaker-specific acoustic patterns (e.g. some may not be perceivable)
  - Which ones do they use?
  - Goal of this study: test the hypothesis that a speaker's utterance-final glottalization pattern is one of these features

Utterance-final glottalization

- Is utterance-final glottalization one of the potential features listeners use to recognize familiar voices?
  - Definition of glottalization: intermittent perceivably-irregular vocal fold vibration; spans a longer time utterance-finally
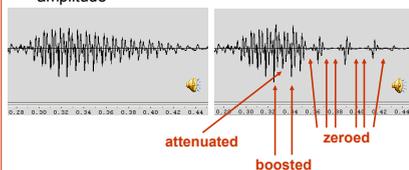

modal   **glottalized**
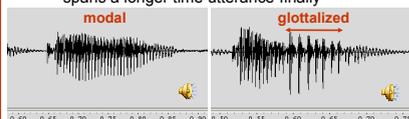
- Rate of occurrence varies substantially among speakers, especially at the ends of utterances (Slifka 2006; Redi, Shattuck-Hufnagel 2001; Henton, Bladon 1987)
- In a perceptual experiment we tested if this is a feature listeners use to recognize familiar voices

Perceptual experiment

- Speakers: frequent and rare 'glottalizers'
- Listeners: familiar with the speakers' voices
- Stimuli: pairs with different final voice quality (VQ)
- Paired comparison tests: which token was the speaker's voice?
- Prediction: listeners will choose the one with the speakers' usual final voice quality as that speakers' voice

Voice quality manipulation

- modal → glottalized: some pitch periods scaled to zero, others scaled to attenuate or boost the amplitude

attenuated   boosted   zeroed
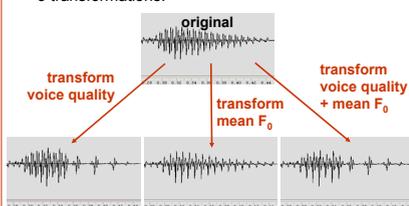
- glottalized → modal: irregular portion replaced by a modal ending and smoothed to the preceding region
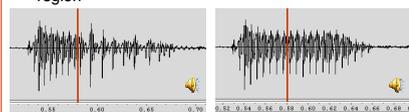
- Naturalness and roughness of the manipulated stimuli was verified in a separate experiment
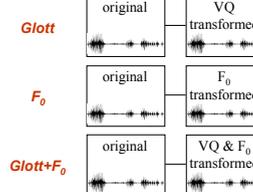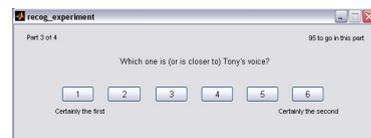
Stimuli

- 3 transformations:

original

transform voice quality   transform mean F_0 transform voice quality + mean F_0

- Why did we also vary mean F_0?
  - to compare the effects of the different manipulations
  - to test the appropriateness of the experimental method

Procedure

- Familiarity test: assessed listeners' familiarity with the speakers' voices
- Paired comparisons: 3 stimulus conditions (pair types):

- The pairs differed only in final VQ, F_0 or both

  Glott: original | VQ transformed
  F_0: original | F_0 transformed
  Glott+F_0: original | VQ & F_0 transformed

- Each pair was tested in both orders

recog_experiment

Part 3 of 4   95 to go in this part

Which one is (or is closer to) Tony's voice?

1   2   3   4   5   6

Certainly the first   Certainly the second

Results: overall

- Familiarity test: recognition rates were well above chance for each listener
- Paired comparisons: calculated the percentage of responses in which the original unmanipulated member of the pair was chosen
- *Glott* condition: if final VQ does not help in recognizing the speaker: listeners would randomly choose between the original and the modified version; results would be ~50% (chance level)

Glot 63%   F0 85%   Glot+F0 90%
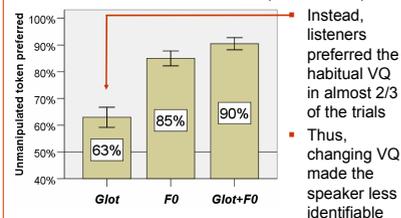
Instead, listeners preferred the habitual VQ in almost 2/3 of the trials

Thus, changing VQ made the speaker less identifiable

Results: by participants

Condition: Glot, F0, Glot+F0

Listener: 1–10   Speaker: FN, MN, FG, MG
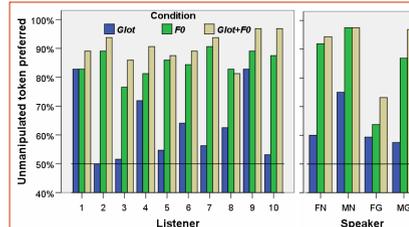
Summary

- Habitual utterance-final voice quality is one of the features listeners use to recognize at least some speakers.
- Results support the idea of a varying feature set depending on the speaker and the listener.
- Familiarity requirement restricts potential listeners
  → subsequent experiment: within-experiment perceptual learning of the speakers' voices
  → very similar results   (Bőhm, Shattuck-Hufnagel 2007)
- How exactly might a featural model work? Can one imagine a non-featural (e.g. Gestalt-based) model?

Acknowledgements

References

Bőhm, T., Shattuck-Hufnagel, S. (2007) "Listeners recognize speakers' habitual utterance-final voice quality", Proc. ParaLing'07, Saarbrücken, pp. 29-34.

Henton, C.G., Bladon, A. (1987) "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, pp. 3-29.

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

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

Van Lancker, D., Kreiman, J., Emmorey, K. (1985) "Familiar voice recognition: patterns and parameters; Part I", J. Phonetics 13:19-38.