Do Listeners Store in Memory a Speaker's Habitual Utterance-Final Phonation Type?

Tamás Bóhm^a Stefanie Shattuck-Hufnagel^b

^aDepartment of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary; ^bSpeech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Mass., USA

Abstract

Earlier studies report systematic differences across speakers in the occurrence of utterance-final irregular phonation; the work reported here investigated whether human listeners remember this speaker-specific information and can access it when necessary (a prerequisite for using this cue in speaker recognition). Listeners personally familiar with the voices of the speakers were presented with pairs of speech samples: one with the original and the other with transformed final phonation type. Asked to select the member of the pair that was closer to the talker's voice, most listeners tended to choose the unmanipulated token (even though they judged them to sound essentially equally natural). This suggests that utterance-final pitch period irregularity is part of the mental representation of individual speaker voices, although this may depend on the individual speaker and listener to some extent.

Copyright © 2009 S. Karger AG, Basel

Introduction

Human listeners often have the experience of recognizing a speaker that they know well, even from hearing just a small portion of a spoken utterance [Pollack et al., 1954]. How listeners do this is a topic of much interest [for reviews see Bricker and Pruzansky, 1976; Van Lancker et al., 1985a; Belin et al., 2004; Sidtis and Kreiman, 2008], and explicit modelling of this cognitive function is a highly challenging task. As a start to addressing this task, the recognition function can be decomposed into two components: the mental representations of familiar speakers [Carterette and Barnebey, 1975], and the process of matching these stored codes to incoming auditory information. (A similar distinction is made in the face recognition model proposed by Bruce and Young [1986].) It is reasonable to assume that, if a voice cue contributes to speaker recognition, then it needs to be part of the first component, i.e. the mental code describing the voice. Note that the existence of such a representation is a necessary but not sufficient condition for

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2009 S. Karger AG, Basel
0031-8388/09/0663-0150
\$26.00/0
Accessible online at:
www.karger.com/pho

Tamás Bóhm
Magyar Tudósok Krt. 2. 1B214
HU-1117 Budapest (Hungary)
Tel. +36 1 463 3512, Fax +36 1 463 3107
E-Mail bohmt@mit.bme.hu

Fig. 1. Example waveform (a) and spectrogram (b) of utterance-final irregular phonation ('(your pa)per' uttered by a female speaker). The arrow marks the irregular region.

its effectiveness in speaker recognition. That is, in order to become useful, this information also needs to play an active role during the matching process. This article addresses the question of whether a familiar speaker's typical utterance-final phonation type is included in the listener's stored representation of that voice.

In considering the issue of voice cues, it is important to distinguish between the set of acoustic differences between speakers that can be reliably measured, on the one hand, and the information that listeners actually store and make use of, on the other. With regard to the first point, there is no doubt that there are a number of acoustic differences among speakers. Such cross-speaker differences may arise both from physiological factors and from learned (habitual, cultural) factors. However, the fact that an acoustic parameter systematically distinguishes one speaker's voice from another does not necessarily mean that it is stored and used by listeners to achieve recognition. A speaker-specific parameter might be useless to the listener if, for example, it cannot be perceived by the human auditory system, or if it is not distinctive across the set of voices known by that listener. Thus one cannot answer the question of which acoustic voice parameters a listener extracts from the speech signal in recognizing the speaker's voice, simply by exploring a speaker's characteristic voice parameters and comparing these parameters across speakers. Instead, once a reliable speaker distinction is established, it is necessary to establish further that listeners make use of that distinction in recognizing familiar voices, i.e. it is both encoded in memory and employed by the matching process. In this study we ask whether one particular voice parameter, i.e. a speaker's habit of producing irregular phonation at the ends of utterances, is part of the implicit mental representation of familiar speakers. Utterance-final irregularity, an example of which is shown in figure 1, is one of several types of intermittent, position-specific variation in phonation type that have been observed to vary across speakers [Dilley et al., 1996].

Before we can ask whether this speaker characteristic is encoded in long-term implicit memory for voices, we need to clarify the term ‘irregular phonation’. For our purposes we have adopted Surana and Slifka’s [2006] explanation:

A region of phonation is an example of irregular phonation if the speech waveform displays either an unusual difference in time or amplitude over adjacent pitch periods that exceeds the small-scale jitter and shimmer differences, or an unusually wide-spacing of the glottal pulses compared to their spacing in the local environment, indicating an anomaly with respect to the usual, quasi-periodic behavior of the vocal folds.

This explanation focuses on irregularity that arises from the vibration of the vocal folds, thus excluding the irregular noise source at the glottis which results in breathy voice [Titze, 1995]. We interpret ‘unusual’ both acoustically and auditorily. That is, for a token to be accepted as irregular phonation, it must exhibit both a visible change from modal (regular) phonation in the waveform/spectrogram, and an audible change in voice quality [Dilley et al., 1996]. Following general practice in the literature on phonation types, we do not set up a particular quantitative criterion for irregular phonation. This is because there are a number of acoustical correlates of irregular phonation (changes in jitter, shimmer, fundamental frequency (F_0), amplitude, open quotient and spectral tilt, etc.) and it is not yet clear what combinations of these parameters define this phonation type. In fact, irregular phonation is an umbrella term: it covers a number of phonation types and phenomena in Titze’s [1995] nomenclature, such as pulsed phonation, creaky voice, glottalized voice and rough voice, as well as diplophonia, aperiodicity and period doubling. It includes most of the categories of vocal irregularity that have been set up by investigators such as Batliner et al. [1993], Hedelin and Huber [1990], and Redi and Shattuck-Hufnagel [2001].

In this study we focus on *utterance-final* irregular phonation as a potentially useful speaker-specific voice characteristic for two reasons. First, informal observations suggest that irregularity in this position has a considerable duration, often lasting for a syllable or more, so it may be more salient to the listener than irregularity in other locations, which often spans a smaller region. Second, the rate of occurrence of utterance-final irregular phonation in English can differ substantially across speakers. For example, Slifka [2006] reports that for 4 speakers of American English, the occurrence of irregular pitch periods in vowel-terminal utterances was 0, 51, 85 and 85%. Slifka [2000, p. 103] noted that each of the 4 speakers in her study appeared to have ‘certain habits to terminate voicing’. Henton and Bladon [1987] and Redi and Shattuck-Hufnagel [2001] also report (for British English and American English, respectively) substantial interspeaker variation in this characteristic.

Based on these earlier studies, it appears that speakers differ in their likelihood of producing irregular pitch periods in utterance-final syllables. However, it is not yet clear whether listeners remember this variation in a particular speaker’s voice. The present study was designed to test the hypothesis that a speaker’s habitual utterance-final phonation type, comprising regular or irregular pitch periods, is part of the listener’s stored knowledge of the voice. In experiment I we compared three alternative ways to compile the stimulus set: original recordings, formant synthesis and manipulated recordings. By ‘manipulated’ we mean that if the end of the original recording was produced with regular phonation, it was amended by a waveform manipulation method to be irregular, and if it was produced with irregular phonation, it was amended to be more regular. In experiment II, we evaluated our potential listeners to determine how well they could recognize the speakers who were familiar to them, testing whether

they have sufficient long-term memories of these voices. Having established their ability to recognize these familiar speakers' voices, in experiment III we used a paired comparison test to determine whether these same listeners store the parameter of utterance-final irregularity for a familiar speaker's voice.

Experiment I

When constructing the stimuli for experiments on the perception of phonation types, the experimenter can choose from a number of methods. For example, natural speech recordings can be used in order to make sure that the stimuli have the same acoustic properties as the phenomenon that is being studied. However, with this method it is impossible to fully control all other aspects of speech. Thus, there may be differences in pitch, loudness or any other parameter among the members of the stimulus set, and these differences can bias the results. This problem can be overcome by using formant synthetic stimuli, a technique that allows full control over the acoustic parameters of the speech samples. This may be the reason why most previous work on the perception of phonation types applied this technique [Klatt and Klatt, 1990; Childers and Lee, 1991; Bangayan et al., 1997; Pierrehumbert and Frisch, 1997]. But with formant synthesis it is likely that not every aspect of human speech is modeled accurately, especially when different phonation types are to be reproduced, and this may result in decreased perceived naturalness. This can make it hard to generalize the results to natural speech. A third alternative is to use natural speech samples that are manipulated in some way [Van Lancker et al., 1985a, b; Kohler, 2000; Allen and Miller, 2004]. Slight manipulations of the acoustic signal may leave naturalness largely intact, while still providing a means of creating minimal pairs that differ only in the aspect under investigation. Such transformation methods may thus provide a compromise between naturalness and control, particularly if care is taken to determine the perceived naturalness of the manipulated tokens.

The aim of the first experiment was to determine which of the above three methods is the most appropriate for our purposes. To this end, test materials included natural unmanipulated speech, formant-synthesized stimuli and tokens created using a novel phonation type transformation method. We evaluated both the degree of perceived roughness and any degradation in naturalness for all three sets of stimuli. As the perception of irregular phonation is usually described as 'rough voice', measuring perceived roughness allows us to assess whether the recordings (especially formant synthetic and transformed) sound like regular or irregular phonation, respectively. Measuring perceived naturalness independently from roughness allows us to determine whether there was a significant degradation in naturalness, which would signal distortions that could bias the results of our further experiments.

Method

Recordings

Nine American speakers (not including the authors) were recorded uttering 2 tokens each of 8 sentences and 4 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. For each of the 216

utterances, the final portion of the utterance was labeled as irregular or regular by the first author, according to the definition discussed in the 'Introduction'. That is, if a potential irregularity in the utterance-final region was either not clearly visible on the waveform/spectrogram or not clearly perceptible by ear, then it was considered regular. This annotation was checked by the second author; there were disagreements in only 19 of the 216 cases, which were resolved by discussions between the two authors [as in Dilley et al., 1996].

The labels for each utterance (see 'Appendix') were used to calculate the likelihood of utterance-final irregularity for each speaker. As expected, for some of the speakers most utterances ended with irregular pitch periods, for some other speakers most utterance endings were regular, and the occurrence rates of irregular phonation for the remaining talkers were not as extreme. From this set of 9 candidate speakers we selected 4 for the perceptual experiment: 1 male and 1 female speaker whose speech frequently exhibited utterance-final irregular pitch periods (87 and 92%) and 1 male and 1 female speaker whose speech seldom did (0 and 17%). The speakers will be referred to as FI, FR, MI and MR, where the first letter of each abbreviation specifies the gender of the speaker (female or male) and the second letter refers to the speaker's habitual utterance-final phonation type (irregular or regular phonation).

Stimuli

The stimulus set for this experiment contained original, phonation type-transformed and formant synthesized speech. The 'original' (natural, unmanipulated) stimuli consisted of the recordings of four words and short phrases (items 1–4 in the 'Appendix') uttered by the 4 selected speakers. All the words ended with a final sonorant, making it possible to observe whether the final pitch periods were regular or not. An utterance which was labeled unanimously as exhibiting the speakers' most typical final phonation type (irregular for FI and MI, regular for FR and MR) was selected from the two recorded versions (corresponding to the shaded cells in the 'Appendix'). For the 2 speakers who frequently produced utterance-final irregular phonation (FI and MI), all of the four selected recordings ended with irregular pitch periods. For the 2 other speakers (FR and MR), who seldom produced this pattern, all the tokens ended with regular phonation.

In a further recording session, we requested the speakers to utter the same texts with their non-habitual phonation type at the ends (regular for those who habitually close utterances with irregular phonation, and vice versa). The aim of this elicitation was to collect tokens that would allow us to compare the roughness ratings of transformed speech with the ratings of both regularly and irregularly ending original recordings produced by the same speaker. The habitually irregular speakers were able to utter the four text items with regular utterance-final pitch periods, so these recordings were included in the stimulus set. However, the other 2 speakers, who habitually produced modal phonation utterance-finally, were unable to consistently produce irregular phonation on request, so such recordings were not available for use in the experiment. We will refer to all natural regular recordings as *orig_reg*, and to natural irregularly ended recordings as *orig_irreg*. In total, there were 16 *orig_reg* and 8 *orig_irreg* tokens.

To create the second set of stimuli we manipulated the phonation type at the end of each original utterance. If the end of the original token was produced with regular phonation, it was amended to sound irregular (rough) and vice versa. In order to make the last portion of a modal recording sound irregular, some of the pitch periods were zeroed out and some others were either attenuated or boosted. This transformation method is described in detail in Böhm et al. [2008] and is briefly explained here. As a first step, the glottal pulses or glottal closure instants were estimated by Praat [Boersma and Weenink, 2006] and hand-corrected by the first author. Then, each glottal cycle was separated out by applying a Hanning window (spanning two cycles) to the environment of each glottal pulse. This procedure is the same as the analysis stage of the PSOLA algorithm [Moulines and Charpentier, 1990] and it extracts the individual cycles into separate waveforms. Note that these waveforms are only approximations of the actual 'cycles' as the impulse responses may overlap in the original speech signal. Each of these one-cycle waveforms (i.e. effectively containing one impulse response each while having a duration of two fundamental periods) was then multiplied by a hand-selected scaling factor (s) and overlapped-and-added to re-synthesize the signal. The scaling factors could either boost ($s > 1$), attenuate ($s < 1$), remove ($s = 0$) or leave unmodified ($s = 1$) an individual cycle (fig. 2).

Fig. 2. Illustration of the transformation method that introduces irregular pitch periods into the speech signal.

When setting the scaling factors, we modeled them after a sample of natural speech with irregular pitch periods (produced by the same speaker, where possible). We attempted to match the irregular pulse pattern (i.e. the pulse spacings and amplitudes) of that sample. For example, when a glottal cycle was substantially longer in the irregular recording than in the modal one, we zeroed out one or two cycles at the corresponding location in the regular waveform. Since a naturally occurring irregular cycle length is not always an integer multiple of the corresponding regular cycle length, this method of period removal usually cannot match the exact length of the irregular cycles, but we believe this is not critical perceptually. As indicated by the results reported below, it seems that the abrupt, substantial cycle length changes introduced by the transformation are sufficient to achieve a rough-sounding voice quality, while the exact lengths of the cycles are apparently of less importance. The relative amplitudes of the irregular pulses in the sample were also copied, in order to mimic amplitude irregularities. The transformation was iteratively fine-tuned by removing some cycles and adjusting the scaling factors until the authors judged it to be both natural-sounding and a perceptually salient example of utterance-final irregularity. The recording with a modal ending in figure 3a was transformed with this method, resulting in the waveform in figure 3b.

To perform the opposite manipulation, i.e. to transform an irregular ending into a modal one, the irregular portion was replaced by a modal ending taken from another utterance (the transformed version of the recording in figure 4a is shown in figure 4b). In 2 of the 8 cases there was no such recording available for the speaker; for these 2 tokens, some groups of pitch periods were copied from the immediately preceding region of regular phonation. None of the copied groups were identical, so that we did not create a strictly periodic signal. The F_0 and amplitude curves of the manipulated endings were then shifted up or down to connect smoothly with the preceding regions.

The 16 originally regular stimuli that were transformed to have irregular endings are denoted as *trans_irreg*. The term *trans_reg* is used for the 8 tokens where the opposite manipulation was employed, to transform irregular tokens into regular ones.

To complete the third set of stimuli for experiment I, we copy-synthesized two versions of each word or phrase produced by the 4 speakers, using the Klatt synthesizer [Klatt and Klatt, 1990]. One version was synthesized with modal and another with irregular phonation in their final regions (*syn_reg* and *syn_irreg*). There were 16 *syn_reg* and 16 *syn_irreg* stimuli in total. By copy synthesis we mean that the time courses of the acoustic parameters that drive the synthesizer were copied from the corresponding natural utterances. While some parameters can be accurately measured on the acoustic signal and fed into the synthesizer, others need to be set using a trial-and-error procedure. We followed the synthesis process described by Klatt and Klatt [1990], by Bangayan et al. [1997] and by Pierrehumbert

Fig. 3. Example of the regular-to-irregular transformation, showing only the last 0.3 s of each utterance. Arrows mark irregular regions. Modal phonation at the end of the recording in **a** was transformed to irregular phonation by removing/scaling selected pitch periods (**b**). When selecting the scaling factors, the utterance in **c** was used as a model.

Fig. 4. Example of the irregular-to-regular transformation, showing only the last 0.3 s of each utterance. The utterance-final irregularity of the recording in **a**, marked by the arrow, was replaced by a regular ending in **b**, that was cut out from the utterance in **c**. Before the concatenation, the surrogate ending's F_0 was adjusted to smoothly join the pitch contour of the 'host' recording.

and Frisch [1997]. Irregular phonation was synthesized by specifying sudden, substantial changes in the F_0 parameter of the synthesizer. In some cases it was necessary to increase the value of additional parameters, including the parameter that controls B1 (first formant bandwidth) during the open portion of the period (DB1), the parameter that makes alternate periods shorter and weaker (DI) and the parameter that introduces random fluctuation in F_0 (FL). Occasionally, an increase in spectral tilt (TL) and in open quotient (OQ) helped to achieve the desired voice quality. Following common practice, male speech was synthesized at a 10-kHz sampling rate and female speech at 13 kHz. Some examples of the synthesizer parameter files are available as an online supplement at www.karger.com/doi/000235658.

Table 1. The stimulus types rated by the listeners in experiment I

Stimulus type	Generation method	Final phonation type	Token count
<i>orig_reg</i>	original recording	regular	16
<i>orig_irreg</i>	original recording	irregular	8
<i>trans_reg</i>	transformed	regular	8
<i>trans_irreg</i>	transformed	irregular	16
<i>syn_reg</i>	formant synthesis	regular	16
<i>syn_irreg</i>	formant synthesis	irregular	16

The six stimulus types, i.e. an original, a transformed and a synthesized version, each with regular and irregular phonation, are summarized in table 1.

Procedure

The experiment consisted of two tests. In one, listeners rated the naturalness of the speech samples, while in the other they judged the roughness. For both tests, responses were given on a 5-point scale. The endpoints of the roughness scale were labeled as ‘not rough at all’ (1) and ‘very rough’ (5), while the extremes of the naturalness scale were denoted as ‘very unnatural’ (1) and ‘very natural’ (5). Listeners heard each token only once and were instructed to respond quickly. Before starting the naturalness test, the entire stimulus set was played for the listeners to demonstrate the range of naturalness that they would encounter during the test. Before the roughness test, listeners heard some examples of natural speech both with and without irregular pitch periods, to clarify the meaning of the term ‘roughness’.

All 80 stimuli were rated twice in both tests, resulting in 160 trials/test for each listener. Presentation order was rerandomized for each listener and each test, with the order of the two tests counterbalanced across listeners. An initial pilot version of the experiment was run with 5 listeners, and the main experiment with 13 listeners (9 males, 4 females, aged 26–68); all were native speakers of American English who were not familiar with the talkers’ voices. (The authors were excluded from the set of listeners in all three experiments reported here.)

In each of the three experiments, audio stimuli were set to equal rms intensity to minimize loudness differences. The experiments were administered via a graphical program (written in Matlab 7.1), using a laptop computer in a quiet office, and stimuli were presented over either AKG K271 or Bose TriPort II headphones. Listeners responded by clicking on the appropriate button displayed on the screen, using the mouse. There was no time limit for responding, but listeners were instructed to respond quickly. In all our analyses, we adopt the significance criterion of $p < 0.05$, except where noted.

Results

After running the experiment with the first 5 pilot subjects, we determined whether listeners could tell which of the original (untransformed) recordings had a rough voice quality, i.e. whether they rated the irregularly terminated recordings as more rough on average than the modal ones. For all 5 subjects this did not appear to be the case; they judged the natural utterances with an irregular ending to have about the same roughness as (or even less roughness than) the natural utterances that had modal endings; differences were in the range of 0.04–0.56 scale point. This surprising result may have arisen because the listeners were unable to perceive the contrast in phonation type in this experimental context, or because they misunderstood the task, mistaking ‘rough’ for ‘unnatural’. Whatever the explanation, if listeners using this method cannot accurately perceive irregularity in unmanipulated stimuli, we cannot expect them to provide

Fig. 5. Mean roughness and naturalness ratings of the six stimulus types. Error bars represent 95% confidence intervals.

an accurate measure of roughness and naturalness of the manipulated stimuli. For this reason, we considered these sessions to be a pilot, from which we learned the importance of drawing the listeners' attention more explicitly to what we meant by 'rough'. In further sessions involving 13 new listeners, we attempted to clarify the term 'roughness' during a short conversation before the experiment: we mentioned famous people who have either noticeably rough or noticeably smooth voices, and pointed out that the phenomenon of rough voice is likely to occur at the ends of utterances. Drawing listeners' attention to the ends of utterances was appropriate, because our aim in this experiment was not to determine whether they naturally pay attention to utterance-final phonation type, but rather to determine how rough or natural this phonation type sounded to them in the formant-synthetic and manipulated stimuli, compared to the natural stimuli. These additional instructions were apparently successful: the 13 listeners who received them rated the natural irregular stimuli to be more than one full scale point (1.16) rougher on average than the natural regular tokens.

For these 13 listeners, a one-way univariate analysis of variance (ANOVA) for naturalness ratings and a separate one for the roughness ratings both showed a significant effect of stimulus type [$F(5,2074) = 489.414$; $p < 0.0005$ and $F(5,2074) = 103.025$; $p < 0.0005$, respectively]. Average scores for each stimulus type can be seen in figure 5 and were further analyzed by Tukey's post hoc tests.

These results show that listeners perceived the difference between utterances with irregular and regular phonation, while accepting both as natural. That is, the natural irregular recordings (*orig_irreg*) received significantly higher ($p < 0.0005$) roughness

ratings than the natural regular ones (*orig_reg*). The naturalness scores of both of these stimulus types were high, and there was no significant difference between them ($p = 0.428$, not significant).

When regular utterance-final phonation was transformed to irregular voice (*orig_reg* → *trans_irreg*) the roughness ratings of the originally modal samples increased by 1.25 point on the 5-point scale ($p < 0.0005$), but this transformation caused only a nonsignificant, 0.20 point decrease in naturalness scores ($p = 0.073$, NS). Not only did the transformation substantially increase roughness, but this increased roughness closely matches the roughness of natural irregularly phonated speech (*orig_irreg*): the difference in the mean ratings was only 0.09 points ($p = 0.957$, NS). Thus, listeners perceived the increased roughness of the transformed utterances, considered them natural, and heard virtually no difference in the degree of roughness of the originally irregular and transformed irregular stimuli.

The reverse transformation of irregular to regular utterance endings was at least partially effective. Perceived roughness decreased by 0.52 of a point ($p < 0.0005$) when natural utterances with irregular phonation were transformed into modal voice (*orig_irreg* → *trans_reg*). However, this perceived roughness was still 0.64 point higher than that of the *orig_reg* stimuli, indicating some remaining degree of nonmodal phonation. This may be due to an audible discontinuity at the concatenation point, or to a slightly rough quality present in regions preceding the concatenation. The associated degradation in naturalness was small (0.07) and nonsignificant ($p = 0.981$, NS).

The naturalness of the formant synthetic stimuli was rated more than two points lower than that of natural or converted speech ($p < 0.0005$ in all pairwise comparisons). These synthetic utterances, whether with regular or irregular endings, were judged to be much rougher than almost all the other samples ($p < 0.0005$ in all pairwise comparisons, except for the synthetic regular vs. original irregular comparison, for which $p = 0.163$, NS). This result suggests either that synthetic speech is inherently perceived as rougher than natural speech (e.g. because of the somewhat lower sampling frequency) or that listeners could not separate the roughness and naturalness aspects of these stimuli (although this did not seem to be the case for the other kinds of stimuli where the two measures are not well correlated). The tokens synthesized with an irregular ending were perceived to be significantly rougher than those synthesized with a regular ending by 0.40 of a scale point ($p < 0.0005$).

Discussion

Taken together, these results suggest that using the transformed stimuli for testing our main hypothesis is appropriate; this method provides a favorable compromise between naturalness and control. Even though we applied a different manipulation to transform regular utterance endings to irregular and vice versa, the transformed speech created by both of these manipulations approached natural irregular and regular utterance endings (respectively) in terms of perceived roughness, with no significant degradation in naturalness. It is thus unlikely that listener responses would be biased by any artifacts caused by the two phonation type transformation methods. Formant synthesis provides a significantly lower level of naturalness, while using original speech samples may introduce confounding factors. Thus the stimulus sets used in experiment III, our main experiment, included tokens manipulated by the method described above.

Experiment II

Before we could test our hypothesis that listeners store information about the utterance-final phonation type of familiar speakers in experiment III, we needed to recruit listeners familiar with the voices of the 4 selected speakers. The aim of experiment II was to quantitatively assess the potential listeners' self-reported familiarity with the voices.

Method

Stimuli

The recordings made for experiment I were adopted: four words and short phrases uttered by 4 speakers (2 males and 2 females). But in this experiment, only the recordings with the speakers' habitual utterance-final phonation type (irregular for FI and MI, regular for FR and MR) were used. From the two available recordings of these words, we used the one that was excluded from experiment I (corresponding to cells in a dark frame in the 'Appendix'). The recordings with the elicited nonhabitual phonation type were excluded. Recordings of the same four words/phrases uttered by a male and a female talker unknown to the listeners were added to the stimulus set as foils, giving 6 possible speakers. The occurrence rate of irregularly phonated utterance endings was high for the male unfamiliar speaker (75% of all tokens) and intermediate for the female unfamiliar voice (37%).

Listeners

Only listeners who were personally familiar with all the 4 speakers' voices participated in the experiment. This requirement severely limited the number of potential listeners – a common problem in studies of familiar speaker recognition [for a discussion, see Van Lancker et al., 1985a]. The 10 listeners (4 females, 6 males) were all faculty members or graduate students at the department where the 4 speakers were affiliated. The listeners were either native speakers of English or had been living in an English-speaking country for at least 3 years. None of these listeners had participated in experiment I.

Procedure

After hearing a token, listeners were asked to select the speaker from the list of 6 speakers (designated by the first names of the 4 known speakers, and 'other male', 'other female'). Each of the 24 recordings (4 words produced by 6 speakers each) was tested twice in different randomized order for each listener.

Results

In order to perform their task (i.e. to explicitly identify the speaker), listeners needed to evaluate three aspects of the voices heard: (a) the gender of the person speaking, (b) the familiarity of the person (if it is a familiar person or not) and (c) the identity and the name of the person. In the face perception literature, face recognition (the sense of familiarity) and person recognition (access to identity-specific semantic information) are usually distinguished [Bruce and Young, 1986]. Based on both the parallels between face and voice processing [Belin et al., 2004] and on everyday observations (occasionally struggling with placing a familiar voice), it is reasonable to assume that these two are separate processes in the auditory modality, too, and thus (b) is a separate aspect from (c).

The results are analyzed in terms of these three aspects. Although our listeners gave a single response, it is straightforward to disentangle the three separate decisions.

For example, when the voice sample was of speaker FR (one of the females well known by the listeners) and the listener chose the first name of speaker FI (the other known female), gender recognition was correct as well as the decision about familiarity, however the identity of the speaker was not established properly. In this experiment, each of these decisions is binary, implying chance levels of 50%.

Gender recognition was perfect: each of the listeners could tell whether they heard a male or a female speaker in all the trials. In 75% of the trials, they could also tell if it was a familiar or an unfamiliar person. For the 10 listeners, the proportion of correct familiarity decisions ranged from 67 to 85%, each significantly higher than chance [one-sample *t* tests; $t(47) \geq 2.424$; $p \leq 0.019$].

When a familiar person was recognized as familiar, the identity of the speaker was established correctly for 90% of the tokens. (Speaker identity decisions for cases when the familiarity decision was mistaken do not carry relevant information.) For 9 of the 10 listeners, one-sample *t* tests showed that speaker recognition rates (75–100%) were significantly higher than chance [$t(31) \geq 3.215$; $p \leq 0.003$]. For the remaining listener the recognition rate of 65% was still above chance although not significantly [$t(30) = 1.662$; $p = 0.107$; NS]. The confusion matrix (table 2) shows that the recognition rates for the 2 familiar female speakers were somewhat lower than for the 2 familiar males. This suggests that the female speakers' voices are harder to recognize, or are more confusable with each other. No clear effect of individual sentence on the confusion patterns was observed.

Discussion

Nine out of our 10 listeners performed well above chance when deciding about the gender, familiarity and identity of the speakers. Thus they can be considered to be familiar with the speakers. For the 1 remaining listener, we could not verify the familiarity with the voices (his performance in identifying the speakers was not significantly higher than the chance level).

Experiment III

In the third experiment, we tested our hypothesis that listeners remember familiar speakers' utterance-final habitual phonation type (in our case, regular or irregular) and can access this information.

In creating the stimuli for experiment III, our aim was to create minimal pairs that differed only in their utterance-final phonation type (i.e. modal vs. irregular) while still sounding natural. In addition, to gain some idea of the size of the effect of irregular phonation, we crossed variation in utterance-final irregularity with variation in the mean F_0 of the utterance. Mean F_0 has been considered to be a parameter that is characteristic of a speaker [for a summary, see Nolan, 1983, p. 124] and that is efficiently employed by humans in speaker recognition [Matsumoto et al., 1973; Abberton and Fourcin, 1978]. Thus it is part of listeners' representations of familiar voices, and is easily accessible during recognition tasks. Varying such a robust cue along with the hypothesized cue of utterance-final phonation type allowed us to compare the relative size of the effect of utterance-final irregular phonation. The F_0 variation condition also

Table 2. Confusion matrix of the speaker identification responses for tokens which were produced by a familiar person and were perceived as familiar

Speaker	Response			
	FR	FI	MR	MI
FR	83%	17%		
FI	20%	80%		
MR			96%	4%
MI				100%

Each row shows the distribution of the responses to a speaker’s utterances. Italicized numbers in the diagonal of the matrix represent correct response rates.

served as a control for the appropriateness of our experimental method: if the method shows an effect of varying mean F_0 , then we can expect it to show a similar effect of varying utterance-final irregular phonation, if this parameter is also stored.

Method

Stimuli

As described in the ‘Methods’ section of experiment I, we selected 4 speakers from a pool of 9: a male and a female who frequently produced utterance-final irregular pitch periods (denoted by MI and FI), and another male and female who seldom did this (MR and FR). The same original, unmanipulated recordings were employed here as in experiment I: four short words and phrases uttered by each of the 4 speakers. All of these utterances exhibited the speakers’ habitual utterance-final phonation type (shaded cells in the ‘Appendix’); the utterances with elicited nonhabitual phonation type were not used.

Each set of stimuli consisted of one such original word or phrase and three manipulated versions of that word. There were 16 such sets, 4 for each speaker, making 64 tokens in total. One of the manipulated versions was the phonation type-transformed token of experiment I (i.e. if the final region was irregular, it was changed to regular and vice versa, using the speech manipulation procedures described at experiment I). The other two manipulated versions were created by the following manipulations:

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 the PSOLA algorithm in Praat [Boersma and Weenink, 2006]. For the lower-pitched male and female, the F_0 was shifted up by the same amount. We chose to apply this uniform F_0 shift for both genders because the difference in mean F_0 between the 2 male speakers and also between the 2 females was roughly 30 Hz. PSOLA is generally considered to have negligible artifacts when it is used to implement such slight F_0 changes [Moulines and Laroche, 1995]. The F_0 modification was not applied to irregularly phonated regions.

Phonation Type and Mean F_0 Transformation. To create the fourth token in each set, both of the other two manipulations were carried out: first, final phonation type was altered, and then the pitch contour was shifted.

Thus each stimulus quadruple included the original utterance, the utterance-final phonation type-transformed version, the mean F_0 -transformed version, and the version transformed in both final phonation type and mean F_0 .

Listeners

The listeners were the same as in experiment II, except that the responses of the single listener whose results did not confirm familiarity with the 4 selected speakers’ voices were discarded, making 9 listeners in total.

Procedure

Forty-eight stimulus pairs were constructed from the 64 tokens described in the 'Stimuli' subsection, in the following way. One member of the pair was an unmanipulated original recording and the other one was a manipulated version of that recording. Thus the pairs differed only in utterance-final phonation type, only in mean F_0 , or in both parameters. 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 first 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'. We chose this sort of scale because it allowed listeners to give both a two-alternative forced-choice decision (i.e. choosing the first stimulus by pressing buttons 1–3 or the second by pressing 4–6) and a 3-point confidence measure (high confidence associated with the endpoints, middle confidence with buttons 2 and 5, and low confidence with the central scale points 3 and 4) in a quick and intuitive way.

Each pair was tested 4 times (yielding 192 trials); the unmanipulated recording occurred twice as the first token of the pair and twice as the second. Presentation order was rerandomized for each listener. A short, optional break was included at the middle of the experiment. For each listener, the test was run in the same session as experiment II (a session lasted 22 min on average, including the duration of the optional break), under similar conditions and with the same equipment.

Results

Our main results can be seen in figure 6, 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 our hypothesis that the typical pattern of utterance-final irregularity is a feature that listeners remember about the speaker. We assume that if this hypothesis is false, then manipulating this aspect of the signal would make no difference in the scores. As the test is forced choice, listeners would then choose randomly between the original and the phonation type-manipulated speech sample. Thus the proportion of correct responses for the 'phonation type' condition would be expected to approximate the 50% chance level. But the results showed something quite different: tokens with the original phonation type were preferred over tokens with changed phonation type 64% of the time. A one-sample *t* test showed that this is significantly higher than chance level [$t(575) = 7.122$; $p < 0.0005$], indicating that listeners could tell the speaker's habitual phonation type in many cases.

For the 4 speakers, the preference rates for the original over the phonation type-transformed tokens in the 'phonation type' condition were between 58 and 78% and these were all significantly higher than chance [$t(143) \geq 2.021$; $p \leq 0.045$]. Across listeners, the rate of correct responses ranged from 50 to 83%. Five out of the 9 listeners scored significantly higher than chance [62–83%; $t(63) \geq 2.049$; $p \leq 0.045$].

For the cases where the F_0 contour was shifted up or down for the transformed member of the pair, the preference rate for the unmanipulated member was 86%. Mean F_0 has been shown to be a robust cue to speaker identity [Matsumoto et al., 1973; Abberton and Fourcin, 1978], and the high proportion of correct responses is consistent with the claim that our method indeed tested some aspect of the memory for voices. Comparison with the 64% correct response rate for pairs of original vs. manipulated phonation type shows that recalling a speaker's typical average F_0 is easier than recalling his/her habitual final phonation type. Nevertheless, altering the phonation type did have a significant effect on listeners' decisions. Moreover, manipulating phonation type had an effect somewhat independent of changing F_0 : the significant increase in the

Fig. 6. Preference for the stimuli with the speakers' usual phonation type, mean F_0 , or both (phonation type + F_0) over their manipulated versions. The horizontal line corresponds to the 50% chance level and error bars represent 95% confidence intervals.

rate of correct responses for the 'phonation type + F_0 ' condition (where both phonation type and mean F_0 were changed) compared to the ' F_0 ' condition [$t(1150) = 2.680$; $p = 0.007$] indicates that the effects of the two parameters 'sum up' in some way. This suggests that, even when a voice characteristic as effective as F_0 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.

To test potential differences across speakers and listeners, an ANOVA was conducted on the preference responses, with condition and speaker as fixed factors and listener as a random factor. The significant interaction between condition and speaker [$F(6,1168) = 5.569$; $p < 0.0005$] showed that the two parameters (i.e. mean F_0 and utterance-final phonation type) may be remembered to a different degree for the different voices in this experiment. The significant interaction between condition and listener [$F(16,1668) = 3.215$; $p < 0.0005$] supports an additional claim about sources of variation in this experiment: i.e. the idea that different listeners remember different cues and utilize them in assessing a particular speaker's voice.

The reliability of our results was assessed by two additional analyses. First, because we tested each pair of stimuli 4 times with each listener, we could measure listeners' internal consistency. Analysis using Cronbach's alpha resulted in a value of 0.789, suggesting that each listener gave similar responses to the four presentations of a pair. This makes it likely that the results are repeatable and not due to random guessing. Second, we compared average implicit confidence ratings: these were significantly higher for correct responses than for incorrect responses [$t(1726) = 18.571$; $p < 0.0005$], adding further support to the idea that actual memory recall of the voices played a greater role in listeners' performance than guessing or other factors.

Discussion

Experiment III tested the hypothesis that, when the listener is given the identity of the speaker, he/she can access the characteristics of utterance-final phonation type of that speaker's voice. When presented with two speech samples from the same speaker,

one with regular and the other with irregular phonation at the end, listeners selected the one with the speaker's typical final phonation type in roughly two thirds of the trials. This is significantly more often than chance would predict, showing that at least some listeners correctly remembered the typical utterance-final phonation type for most of the familiar speakers in this experiment. 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 4 speakers, and for 5 out of 9 listeners).

One might wonder whether these results could arise if listeners chose the unmodified member of each stimulus pair as the one closest to the speaker's voice, or if there were audible artifacts introduced by the phonation type and F_0 transformations. However, there are two arguments that render this explanation less plausible. First, according to the results of experiment I, the phonation type transformation methods yield reasonably good approximations of naturally occurring regular and irregular phonation and we can assume the same for the F_0 transformation. Second, a pilot experiment with a design similar to the present experiment, but using formant synthetic stimuli, yielded very similar results [Böhm, 2006]. In this pilot, listeners preferred the speakers' habitual utterance-final phonation type over its transformed counterpart in 63% of the cases. The preference rate for the condition where the F_0 contour was transformed was 78%, and it was 80% when both phonation type and F_0 were manipulated. These results were obtained for pairs of stimuli with no naturalness differences between the two members, since all of the stimuli were created using formant synthesis. Based on these observations, it is unlikely that the results reported here were substantially affected by any potential stimulus artifacts or unnaturalness.

General Discussion

Because some speakers of English frequently produce intermittent episodes of irregular phonation, and the likelihood that this will occur can vary substantially from one speaker to another, it is possible that regions of irregular pitch periods in certain locations may be one of the acoustic parameters that listeners encode in memory about a speaker and use in the process of recognizing that speaker. This study examined the former possibility, i.e. whether the speaker's typical patterns of pitch period irregularity at the end of utterances are remembered. Results support this hypothesis by showing that listeners encode in memory information about whether the talker's likelihood of producing irregular pitch periods utterance-finally is high or low, and can implicitly access this information. That is, when asked to select which of a pair of speech samples was closer to a familiar speaker's voice, most listeners tended to choose the sample that had the speaker's typical utterance-final phonation type, instead of choosing the transformed sample which had a different final phonation type.

The type of intermittent irregular phonation that was the subject of this study spans only a limited portion of the utterance duration (even though it is often longer than in other locations, such as phrase-initially or at pitch accents). Despite this relatively limited time span, our results show that listeners store this information and can employ it in this task. The evidence for this claim is that utterance-final irregular phonation often helped listeners, both by itself (i.e. when this parameter alone was manipulated) and

also when a different (and very robust) voice characteristic (i.e. mean F_0) was manipulated at the same time.

This finding provides some information about the ways in which listeners store and use information about familiar speaker voices. According to the results of Van Lancker et al. [1985a], the set of features extracted for speaker recognition may vary with different speakers and different listeners, probably based on their usefulness in distinguishing among a set of known voices. Our results illustrate some aspects of this variation in listener behavior: although the effect of irregular phonation at the ends of utterances was significant for the majority of our speakers and listeners, there were differences among them. It seems that not all of our listeners could remember irregular phonation patterns, and for some speakers this information may not have been distinct enough to be useful (and thus not remembered or accessed). In almost any model of speaker recognition, whether feature-based or Gestalt-based [Siddis and Kreiman, 2008], before we can ask questions about how cues are integrated into a decision about speaker identity (the process of recognition), we need to understand what information is stored about the voice (i.e. what is the repertoire of cues with which the recognition process can work). Our results contribute to the development of speaker recognition models by providing evidence that utterance-final voice characteristics are included in this set of available cues, at least for some speakers and for some listeners.

This study focused on the representation of voices that are personally familiar to the listener. This is convenient, since no training is required, but it also has some disadvantages: requiring potential listeners to be familiar with the voices in the experiment severely restricts the number of possible participants and also limits control of the listeners' degree of familiarity with the test voices. Within-experiment perceptual learning of talkers' voices would enable recruitment from a wider population, and would allow control over the amount of previous exposure to the voices.

Traditionally, studies of memory for voices and of speaker recognition have focused on globally defined voice parameters, such as mean F_0 , mean formant frequencies and speaking rate. Our results suggest, however, that characteristics appearing intermittently during an utterance can also become part of the mental representation of a voice. The findings of Allen and Miller [2004] are also in line with this conclusion: they showed that listeners remember another intermittent speaker characteristic: short vs. long voice onset times for word-initial voiceless stop consonants. In this sense, our results support the conclusion drawn by Pollack et al. [1954] that it is not simply the duration of the speech sample that is important for recognition, but also how well it samples the repertoire of the speaker's voice characteristics. That is, in order to provide enough information for maximum performance in determining a speaker's identity, a sample utterance needs to be long enough to allow the detection of patterns that appear only intermittently in the speaker's output. Finally, these results support the view that related speech technologies, such as automatic speaker identification and voice conversion, may benefit from using such intermittent acoustic patterns. For example, Malyska [2008] successfully used irregularity-based acoustic features in automatic speaker identification.

Appendix

Table 1. Phonation type labels for the final region of the two utterances of the sentences by each of the 9 speakers

	Speaker																	
	F1		MR		FI		M1		FR		F2		M2		MI		M3	
<i>Yesterday</i>	+	+	–	–	+	+	–?	–?	–	–	–	+	–	+	+	+	–?	+
<i>Ready</i>	–	–	–?	–	+	+	+	–	–	–?	–	+	–	–	+	+	–	–?
<i>Your paper</i>	–	+	–	–	+	+	+	+	–	–	–	+	–	–	+	+	–?	–?
<i>Debby</i>	–?	–	–	–	+	+	–	–	–	–	–	+	–	–	+	+	–	–
<i>I saw your paper yesterday.</i>	+	+	–	–	+	+	–	+	–	–	–	+	–	–	+	+	+	+
<i>Yes, you did it.</i>	+	+	–?	–	+	+	–	–?	+	+	–?	+	–	–	+	+	+	+
<i>What you do is up to you.</i>	–	–	–	–	+	+	+	+	–?	–	–	–?	–	–	+	+	–	–
<i>Debby debated about potatoes.</i>	–	–	–	–	+	+	+	+	–	–?	–	–	–	–	+	+	–	+
<i>Yesterday Roy sought a bed.</i>	–	–	–	–	+	+	+	+	–	–	+	+	–	–	+	–?	+	–
<i>What did Debby debate?</i>	+	+	–	–	+	+	+	–	+	–	+	–	–	–	+	+	–	–
<i>Your paper is ready.</i>	–	–	–	–	+	+	–	–	+	–	+	+	–	–	+	+	–	–?
<i>Did you do it yesterday?</i>	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–

The first letter in the identifier of the speakers refers to their gender. The 4 speakers whose voices were used in the experiments are denoted with the same letter codes as in the experiments (FR, FI, MR and MI). + = Irregular phonation; – = regular phonation. A question mark next to a label shows that there was disagreement between the two labelers. Shaded cells denote recordings that were part of the stimuli in experiments I and III, while cells in a dark frame are utterances used in experiment II.

Acknowledgements

The authors are grateful to Kenneth N. Stevens for hosting the first author at the MIT Speech Communication Group where this work was begun, to Janet Slifka for her many helpful insights concerning the design of the experiments, to Kushan K. Surana for the initial idea of experiment III, to Géza Németh for his support, to the Voice Quality Study Group for the inspiring discussion of the manuscript and to two anonymous reviewers for their insightful feedback, and to the Hungarian-American Fulbright Commission for help in recruiting listeners for experiment I. The first author was supported by a Fulbright Scholarship in the initial stage of this work and then by a TELEAUTO-NKFP07 grant, and the second author by NIH grants RO1-DC002978 and RO1-DC0075.

References

- Abberton, E.; Fourcin, A.J.: Intonation and speaker identification. *Lang. Speech* 21: 305–318 (1978).
- Allen, S.J.; Miller, J.L.: Listener sensitivity to individual talker differences in voice-onset-time. *J. acoust. Soc. Am.* 115: 3171–3183 (2004).
- Bangayan, P.; Long, C.; Alwan, A.A.; Kreiman, J.; Gerratt, B.R.: Analysis by synthesis of pathological voices using the Klatt synthesizer. *Speech Commun.* 22: 343–368 (1997).
- Batliner, A.; Burger, S.; Johne, B.; Kiessling, A.: MÜSLI: a classification scheme for laryngealizations. *Proc. ESCA Workshop on Prosody, Lund 1993*, pp. 176–179.
- Belin, P.; Fecteau, S.; Bédard, C.: Thinking the voice: neural correlates of voice perception. *Trends cogn. Sci.* 8: 129–135 (2004).
- Boersma, P.; Weenink, D.: Praat: doing phonetics by computer (version 4.4.10); computer program, retrieved February 22, 2006, from <http://www.praat.org/>
- Böhm, T.: Is utterance-final glottalization a cue for speaker recognition by humans? 151st Meet. ASA, Providence 2006.
- Böhm, T.; Audibert, N.; Shattuck-Hufnagel, S.; Németh, G.; Aubergé, V.: Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. *Proc. Acoustics '08, Paris 2008*, pp. 6141–6146.
- Bricker, P.D.; Pruzansky, S.: Speaker recognition; in Lass, *Contemporary issues in experimental phonetics*, pp. 295–326 (Academic Press, New York 1976).
- Bruce, V.; Young, A.: Understanding face recognition. *Br. J. Psychol.* 77: 305–327 (1986).
- Cartrette, E.C.; Barnebey, A.: Recognition memory for voices; in Cohen, Nooteboom, *Structure and process in speech perception*, pp. 246–265 (Springer, Berlin 1975).
- Childers, D.G.; Lee, C.K.: Vocal quality factors: analysis, synthesis, and perception. *J. acoust. Soc. Am.* 90: 2394–2410 (1991).
- Dilley, L.; Shattuck-Hufnagel, S.; Ostendorf, M.: Glottalization of word-initial vowels as a function of prosodic structure. *J. Phonet.* 24: 423–444 (1996).
- Hedelin, P.; Huber, D.: Pitch period determination of aperiodic speech signals. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Albuquerque 1990*, pp. 361–364.
- Henton, C.G.; Bladon, A.: Creak as a sociophonetic marker; in Hyman, Li, *Language, speech and mind: studies in honour of Victoria A. Fromkin*, pp. 3–29 (Routledge, London 1987).
- Klatt, D.H.; Klatt, L.C.: Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. acoust. Soc. Am.* 87: 820–857 (1990).
- Kohler, K.J.: Investigating unscripted speech: implications for phonetics and phonology. *Phonetica* 57: 85–94 (2000).
- Malyska, N.: Analysis of nonmodal glottal event patterns with application to automatic speaker recognition; PhD diss. Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge (2008).
- Matsumoto, H.; Hiki, S.; Sone, T.; Nimura, T.: Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Trans. Audio Electroacoust.* AU-21: 428–436 (1973).
- Moulines, E.; Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* 9: 453–467 (1990).
- Moulines, E.; Laroche, J.: Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Commun.* 16: 175–205 (1995).
- Nolan, F.: *The phonetic bases of speaker recognition* (Cambridge University Press, Cambridge 1983).
- Pierrehumbert, J.B.; Frisch, S.: Synthesizing allophonic glottalization; in van Santen, *Progress in speech synthesis*, pp. 9–26 (Springer, New York 1997).
- Pollack, I.; Pickett, J.M.; Sumby, W.H.: On the identification of speakers by voice. *J. acoust. Soc. Am.* 26: 403–406 (1954).
- Redi, L.; Shattuck-Hufnagel, S.: Variation in the realization of glottalization in normal speakers. *J. Phonet.* 29: 407–429 (2001).
- Sidtis, D.; Kreiman, J.: Let's face it: phonagnosia happens, and voice recognition is finally familiar; in Pachalska, Weber, *Neuropsychology and philosophy of mind in process – essays in honor of Jason W. Brown* (Ontos, Frankfurt 2008).
- Slifka, J.: Respiratory constraints on speech production at prosodic boundaries; PhD diss. Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge (2000).
- Slifka, J.: Some physiological correlates to regular and irregular phonation at the end of an utterance. *J. Voice* 20: 171–186 (2006).
- Surana, K.; Slifka, J.: Acoustic cues for the classification of regular and irregular phonation. *Proc. 9th Interspeech, Pittsburgh 2006*, pp. 693–696.
- Titze, I.R.: Definitions and nomenclature related to voice quality; in Fujimura, Hirano, *Vocal fold physiology – voice quality control*, pp. 335–342 (Singular, San Diego 1995).
- Van Lancker, D.; Kreiman, J.; Emmorey, K.: Familiar voice recognition: patterns and parameters: part I. *J. Phonet.* 13: 19–38 (1985a).
- Van Lancker, D.; Kreiman, J.; Wickens, T.D.: Familiar voice recognition: patterns and parameters: part II. *J. Phonet.* 13: 39–52 (1985b).