Interspeech 2018
2-6 September 2018, Hyderabad, India

Influences of fundamental oscillation on speaker identification in vocalic utterances by humans and computers

Volker Dellwo[1], *Thayabaran Kathiresan*[1], *Elisa Pellegrino*[1], *Lei He*[1]
Sandra Schwab[1], *Dieter Maurer*[2]

[1]Department of Computational Linguistics, University of Zurich
[2]Department of the Performing Arts, Zurich University of the Arts
volker.dellwo@uzh.ch

Abstract

We tested the influence of fundamental oscillation (*fo*) on human and machine speaker recognition performance in vocalic test utterances. In experiment I, we trained a Gaussian-Mixture model on 15 speakers (80 multi-word utterances each) and tested it with sustained vowel utterances (/a:/, /i:/ and /u:/) under six *fo* conditions, three changing (fall, rise, fall-rise) and three steady-state (high, mid, low). Results revealed better performance for the steady-state compared to the changing conditions and within the steady-state condition, performance was poorest for high *fo*. In experiment II, we tested 9 human listeners on a subset of 4 speakers from experiment I. They went through two training tasks (training 1: multi-word utterances; training 2: words). In the test, they recognized speakers based on the same vocalic utterances as in experiment I (for these 4 speakers). Results showed that performance was about equally high for the changing and steady-state vowels, however, in the steady-state condition performance was best for high *fo* vowels. The experiments suggest that (a) *fo* has an influence on the strength of speaker specific characteristics in vowels and (b) humans - compared to machines - pay attention to different acoustic information in vocalic utterances for speaker recognition.

Index Terms: speaker idiosyncratic information in vowels, automatic speaker recognition, auditory speaker recognition

1. Introduction

Next to information about speech, vocalic utterances contain information about the speakers themselves. For this reason, speaker recognition is to some degree possible based on vocalic utterances alone, either in non-speech vowel-like hesitations or during vocalic parts from the speech signal [2]. Vocalic utterances contain information about (a) the speaker's source signal, i.e. fundamental oscillation (*fo*) characteristics and the quality of vibration (voice quality) and (b) the speaker's vocal tract transfer function. Both these factors have anatomic correlates, i.e. size and mass of the vocal folds and length and diameters of the vocal tract cavities. Variability in the size and mass of the vocal folds results in average *fo* variability between speakers which is one important speaker specific characteristic. The vocal tract transfer function is to some degree dependent on a spectrally dense source signal for individual vocal tract characteristics to be revealed. For example, when *fo* of a speaker is relatively low, the spacing between the harmonics is narrow, resulting in rich detail of the vocal tract. With increasing *fo*, the spacing between harmonic becomes wider and thus the vocal tract transfer function is undersampled. Hence, in steady state vocalic utterances of high *fo* it should be more difficult to retrieve speaker specific vocal tract characteristics. In addition, a high *fo*, which is higher than a speaker's average, contains less information about the speaker's individual source characteristics which, in addition, should lead to a loss in speaker specific characteristics. Given this situation, we hypothesized that speaker recognition in relatively high *fo* utterances should be poorer than in utterances at comparatively low *fo* or in vocalic utterances in which *fo* varies.

We tested this hypothesis with a computer recognition model and with human listeners. In both cases, we used sentence utterances as training material and vocalic utterances as test material. The vocalic utterances were produced by humans with either a low, a mid, or a high steady state *fo* (level tones) or a rising, falling, or fall-rise *fo* (contour tones).

Computers and humans apply fundamentally different techniques in voice recognition. While humans pay significant attention to average *fo* characteristics of the speaker [1, 4, 5, 8], computer models are much more restricted concerning this feature. In the case of an MFCC acoustic feature extraction, the information about *fo* is systematically excluded from the analysis and attention is paid predominantly to the individualities of the vocal tract transfer function. We thus expected that humans would rely to a high degree on the *fo* in a test vowel to be representative of a speaker's average *fo* for a correct recognition. This means that recognition performance in mid-level tone vowels should be higher than in low- or high-level tone vowels. Given that in contour tones the typical range of *fo* is present, performance should be highest. Since *fo* characteristics are only rudimentarily present in MFCCs, a computer model, based on such feature extraction characteristics, should perform equally well in contour tones when vocal tract characteristics are sampled well through the sweeping tone behaviour. Their performance should be particularly poor in high *fo* level tones, as important detail about the vocal tract transfer function is missing.

In experiment I, we trained a Gaussian-Mixture speaker recognition model (GMM) on sentence utterances of 15 speakers and tested it on vocalic utterances with different level and contour tones. The choice of GMM was motivated by the relatively small training data size and number of speakers. As humans cannot be trained easily on such a large number of speakers, we selected a subset of 4 speakers and a subset of the training and test material and tested them in a behavioral speaker recognition task in experiment II.

10.21437/Interspeech.2018-2331

2. Experiment I: computer speaker identification

2.1. Method

Speakers: 15 speakers of Zurich German (11f, 4m) aged between 20 and 30 years. Speakers were all students at Zurich University and received a small payment for their participation.

Recordings: All recordings took place in a sound-treated booth at Zurich University. Each speaker read 80 sentences in Standard German ranging between 9 and 37 syllables. Subsequently, speakers were asked to produce different tone utterances. Speakers were given examples from Chinese tone productions to understand the notion of contour (rising, falling, fall-rise) and level tones (low, mid, high). Speakers were trained in the tone productions by a Chinese native speaker (fourth author). After practicing the tone productions they were then recorded producing three repetitions of the following tones for each of the vowels /a:/, /i:/ and /u:/:

- **level low (lvlo)**: low steady-state pitch
- **level mid (lvmd)**: mid steady-state pitch
- **level high (lvhi)**: high steady-state pitch
- **contour falling (fall)**: falling pitch contour
- **contour fall-rise (fari)**: fall followed by rise
- **contour rising (rise)**: rising pitch contour

The total number of vowel utterances in the test was 810 (15 speakers * 3 vowels * 6 tones * 3 repetitions).

Sound editing: All vowel productions were cut to be precisely 1.5 seconds in duration and were faded-in and faded-out for 10% of this duration each. For the natural vowel production the mid-point in time was identified and the extraction was +/- 0.75 sec around the mid-point. Fig. 1 shows an example of an extracted falling tone in the vowel /a:/. Formants (red dots) were constant while *fo* (blue line) was consecutively falling.

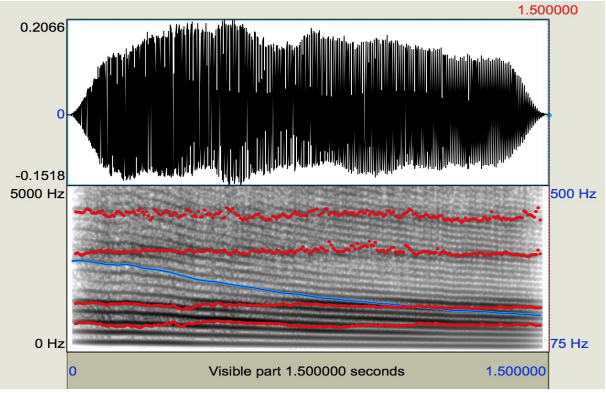

Figure 1: *Waveform (top) and spectrogram (bottom) with superimposed formant (red dots) and pitch estimations (blue dots), showing a vowel /a:/ with falling tone for a female speaker.*

Automatic recognition: MFCCs with 13 coefficients were calculated for all speech material based on 25 ms frames with a frame shift of 10 ms. A speaker identification model based on 10 Gaussian-mixtures with diagonal covariance was used (designed by the second author). The model was trained on the 80 sentence utterances by each speaker and for the test we used the 1.5 sec vowel utterances.

Statistical analyses: Statistical analysis was carried out using R (version 3.1.3; R Development Core Team, 2016; lmerTest R package; [7]). We ran a mixed-effects logistic regression model on the correct/incorrect responses [3]. The fixed part of the model was either comprised of 'level tone' or 'contour tone'. The random part of the model included random intercepts for items (the variable 'item' corresponds to 'Vowel_Speaker'; each vowel produced by each speaker has been repeated 3 times) and random slopes allowing for the effect of 'tone' or 'contour' to differ across items. The significance of the main effect was assessed with likelihood ratio tests that compared the model with the main effect to a model without it. Post-hoc analyses with Tukey correction for multiple comparisons were performed in the case of 'tone' to obtain 2 by 2 comparisons. All statistical analyses were performed on raw data (correct/incorrect responses).

2.2. Results & Discussion

Fig. 2 shows mean identification scores under each of the tones. Descriptive results suggested that speaker identification was better under the contour tones compared to the level tones; this effect was significant ($\chi2(1)=8.88$, $p=.003$), i.e. contour tones were better identified than level tones. Overall we obtained an effect of tone ($\chi2(5)=18.20$, $p=.003$). Post-hoc analyses (with Tukey adjustments) revealed, however, that only the effects for contrasts lvhi-fall ($p=.007$) and lvhi-rise ($p=.03$) were significant. Within the level tones we can see a higher performance for lvmd compared to lvlo and lvhi, however, this effect was not significant within the present framework.

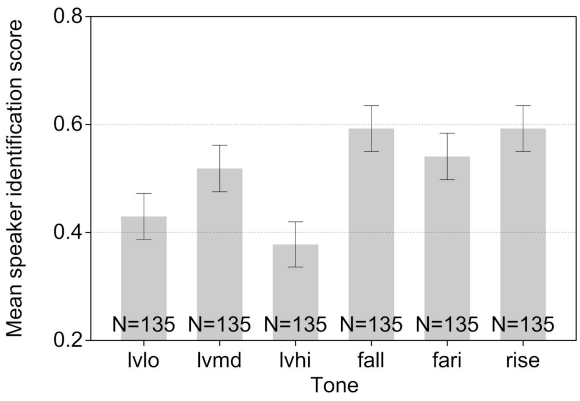

Figure 2: *Mean identification scores (15 speakers * 3 vowels * 3 repetitions = 135) with +-1SE (y-axis) for each of the six tone productions (x-axis). N=135 from Speaker identification was worth under level tones compared to contour tones.*

Fig 3 contains two histograms for *fo* (interquartile range) for the female speakers, one for the training data (top) and one for test data (bottom), with level tones in different colours (contour tones not included). The figure reveals that low tones were produced in the low range of *fo* of the training population, while high tones were produced in a range exceeding the range of the training data. Mid tones were typically produced in a mid to upper range in respect to the training population.

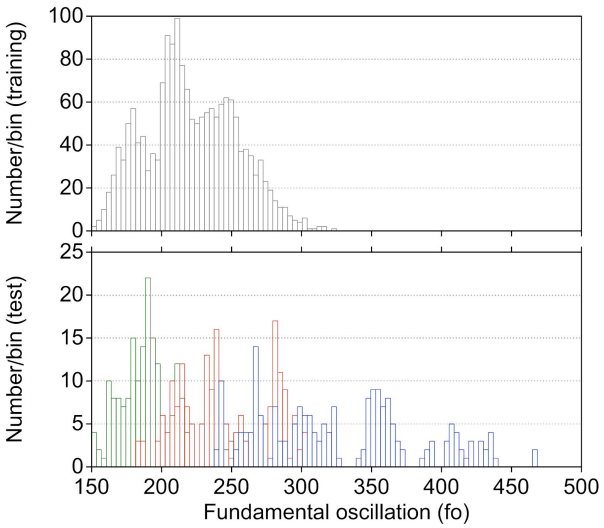

Figure 3: *Histograms of fundamental oscillation (interquartile range) for female speakers in training data (grey), lvlo (green), lvmd (red), lvhi (blue).*

These observations in the data allow some conclusions about the role of *fo* in vocalic utterances in the GMM: (a) The significant performance difference between high level tones and rising and falling contour tones provides evidence for the hypothesis that at high *fo* important detail in the transfer function of the vocal tract is missing which limits the performance of the recognition model. (b) Even though MFCC features are predominantly sampling characteristics about the vocal tract, performance has a tendency to be better when *fo* in the test samples occurs at frequencies similar to the training data. This might either mean that some residual *fo* information is still present in the MFCC data or that some effects on *fo* on the vocal tract characteristics during production are revealed by the MFCCs. (c) It is unclear whether this observation can be generalized to other speaker identification models - most likely not. There are numerous parameter settings that will influence the overall performance of models but it remains unclear whether that has an influence on the relative recognition performance between tone categories, in particular the observed influence of level tones. The observation, however, suggests that also in MFCCs there is more of an influence of *fo* than might be expected. It will be interesting to see how models using iVectors will perform in comparison.

3. Experiment II: human perception

3.1. Method

Listeners: 9 native German listeners from Switzerland. Listeners were students at the university of Zurich between 20 and 30 years of age. Each listener received a small payment for their participation.

Stimuli: From the 15 speakers used in experiment I, 4 female speakers were selected that were auditively well separable for human listeners (speakers 8, 9, 12 and 13). Since we expected that speaker recognition based exclusively on vowels would be a difficult task for listeners, we estimated that four speakers would be an appropriate number. Speakers were selected by the first and second authors based on auditive criteria. The exact same vocalic utterances as in the test of experiment I were used in the present experiment without repetition. The total number of stimuli was 72 (4 speakers * 3 vowels * 6 tones).

Procedure: Listeners were tested with a computer interface programmed in Praat. Listeners' task was to learn to attribute voices of the four female speakers to sketches of females that were well distinguishable by shape and colour on the screen. Fig. 4 shows a screenshot of the basic interface during familiarization phase. Listeners went through the following steps: **Familiarization** with the voices by clicking on any of the characters and then hearing a sample of their voice from a random utterance. **Training 1:** 20 sentences (5 from each speakers) were presented randomly. After each presentation listeners had to choose which female sketch they thought the voice belongs to. Feedback was provided. In case listeners were wrong, they were presented the correct sketch enlarged on the screen and heard the voice again. If listeners had >70% correct they moved to training 2, otherwise they had to repeat training 1 once. **Training 2:** Identical to training 1 but with 20 word utterances (5/speaker). Word utterances were chosen in the second training to direct listeners attention to short-term speaker specific characteristics. This should avoid listeners to pay too much attention to suprasegmental characteristics of speech (e.g. intonation or rhythm) which might not be helpful in the subsequent identification based on vowels. **Test:** One repetition of each vowel and each tone for each of the speakers (3 vowels * 6 tones * 4 speakers = 72 total) was presented and listeners had to choose the respective sketch without feedback.

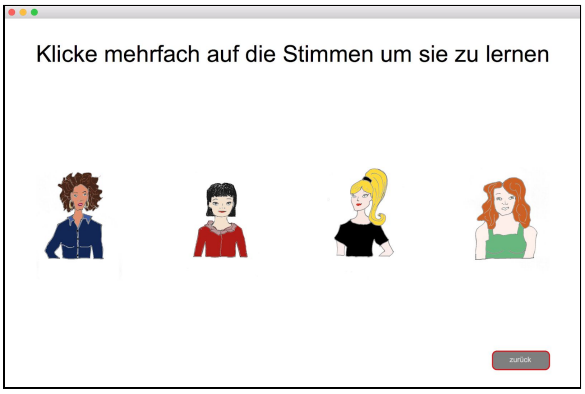

Figure 4: *Screenshot of the interface. Listeners learned to attribute the voices to the fictive characters on the screen.*

Statistical analysis: Statistical analysis was carried out for 'level tone' and 'contour tone' differences using the same mixed-effect logistic regression model as in experiment I. Here we did not include random slope for item since the tone information was already contained in the file. No intercept for listener was included since it did not improve the model.

3.2. Results & Discussion

Results for mean identification scores in human listeners for the different tones are presented in Fig. 5 (top). Descriptively, results show that there is not much of a difference between level and contour tones apart from the high-level tones for which performance was best.

Inferentially, an effect could neither be obtained for contour tone ($\chi 2(1)=.928$, $p=.761$) nor for level tone ($\chi 2(5)=2.8744$, $p=.719$). We argue that the number of listeners is currently too low for effects to be visible. We are currently running more listeners to test whether the present descriptive difference between high level and the other tones will reveal to be significant. One observation supporting the view that listeners might be better at speaker identification for high level tones is that 8 out of 9 listeners showed the same pattern; only one listener performed better in case of the mid level tones (0.33 for high, 0.58 for mid).

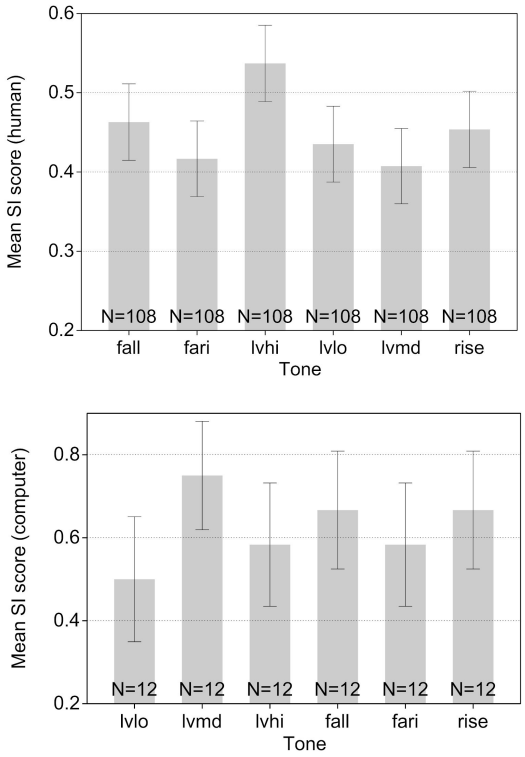

Figure 5: *Mean identification scores (y-axis; 4 speakers * 3 vowels * 9 listeners = 108) for each of the six tone productions (x-axis) for humans (top) and for the exact same data by the identical computer model of experiment I (bottom; 4 speakers * 3 vowels).*

Additional testing: Human recognition results were qualitatively different from computer performance in that humans reveal a tendency to show a higher performance for high level, while computers are better for mid level tones. Also, the computer model performed significantly better for contour compared to level tones. Given that the computer model in experiment I received significantly more speakers, different genders and more training and test items than humans, we ran the computer model on the reduced training and test data used for humans in experiment II. Results are plotted in Fig. 5 (bottom). Since only one model could be run on the 72 stimuli test-set, there were only 12 test items for each tone condition. Given this small number, we did not process inferential test models. While results must be taken with caution, the computer model replicates the higher performance for mid level tones in the level tones obtained in experiment I. The effect of a higher performance under contour tones is not visible in Fig. 5 (bottom). It is thus possible that the contour effect obtained in experiment I might be related to the higher number of speakers and is possibly not present in the 4 speaker sub-choice in experiment II. In other words: If humans were trained and tested on the 15 speakers in experiment I, they might show a contour effect too.

4. Conclusions

In the present experiments we tested the influence of *fo* in vocalic utterances on speaker identification performance by machines and human listeners. We hypothesized that an increasing *fo* should pose difficulties on computer recognition systems to identify speakers and we can confirm that this was the case in our experimental setup. There was a significant difference between high level and contour tones, implying that GMMs rely on dense spectral information to obtain sufficient speaker specific detail in the MFCCs. However, at low *fo* when spectral information is densest, performance was similarly poor as at high level tones. We suppose that this might have to do with the fact that the production of low level tones influences the vocal tract characteristics in a way that is atypical for what can be found in the training data.

We also hypothesized that humans might make more use of *fo* information in the recognition process compared to machines, when MFCCs are used for the acoustic modelling. Surprisingly, humans showed a tendency to be best in speaker recognition at high level tones, when (a) *fo* average is most atypical compared to what they learned in the training voices and (b) spectral information is sparse. It seems implausible that listeners are better at identifying speakers under such conditions. However, the finding might might be in line with recent observations that vocal tract detail in vocalic utterances has been underestimated in vowels produced at high *fo*. [6] showed that vowels produced by soprano voices still contain sufficient information about vowel identity, even though formant information is vastly deteriorated. It will be interesting to study whether speaker specific information is also maintained at higher frequencies in human listener.

The present experiments should be seen as a proof of concept for a setup in which computer and human performance can be compared for speaker recognition in vowels. Drawing direct comparisons between humans and computers is typically difficult as different types of test and training data make the comparisons nonsensical. We argue that more controlled datasets on vocalic variability, larger numbers of listeners for the human testing and different types of computer recognition models will allow us to understand better, which information in vocalic utterances is most relevant for speaker recognition purposes.

ACKNOWLEDGEMENTS: Iuliia Nigmatulina ran the experiments and produced the sketches for the human test interface. Honglin Cao commented on the draft.

5. References

[1] E. Abberton and A. J. Fourcin, "Intonation and speaker identification," *Language and Speech*, vol. 21, pp. 305–318, 1978.

[2] K. Amino and T. Arai, "Contribution of consonants and vowels to the perception of speaker identity," *Japan-China Joint Conference of Acoustics*, (C), CD-ROM. 2007

[3] R. H. Baayen, D. J. Davidson and D. M. Bates, "Mixed effects modeling with crossed random effects for subjects and items," *J. Memory Lan*, vol. 59, pp. 390-412, 2008.
[4] O. Baumann and P. Belin, "Perceptual scaling of voice identity: Common dimensions for different vowels and speakers," *Psychological Research*, vol. 74, no. 1, pp. 110–120, 2009
[5] W. A. V. Dommelen, "The contribution of speech rhythm and pitch to speaker recognition," *Language and Speech*, vol. 30, pp. 325–338, 1987.
[6] D. Friedrichs, D. Maurer and V. Dellwo, "The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz," *The Journal of the Acoustical Society of America*, vol. *138*, no. 1, pp. EL36-EL42, 2015.
[7] A. Kuznetsova, P. B. Brockhoff and R. H. B. Christensen, "lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-20,", 2014.
http://CRAN.R-project.org/ package=lmerTest (Last viewed October 24, 2016).
[8] C. Lariviere, "Contributions of fundamental frequency and formant frequencies to speaker identification.," *Phonetica*, vol. 31, pp. 185–197, 1975.