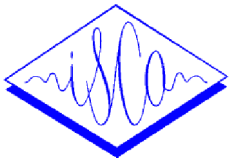

Pinpointing the Difference – Visual Comparison of Non-Native Speaker Groups*

Florian Hönig[1], *Sebastian Wankerl*[1], *Anton Batliner*[1,2], *Elmar Nöth*[1]

[1]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[2]Machine Intelligence & Signal Processing Group, TUM, Munich, Germany
{florian.hoenig}@fau.de

Abstract

We apply a tool originally developed for comparing pathological and healthy speakers to non-native speech. The method works on speakers who produce a given word sequence. Using time-alignment, we can display prototypical loudness contours, local tempo variations, and also spectrograms, together with information on variability and group effect size over time. The system, which will be made publicly available, is able to expose typical differences in a group of German and Italian speakers.
Index Terms: pathological speech, non-native speech, visualization, interpretation, acoustic features

1. Introduction

Characterizations of non-native speech are often available as stereotypes; for a given database, one can listen through the recordings and obtain a subjective impression. However, these are often hard to translate into acoustical correlates needed to design systems for automatic assessment of non-native speech. We show that *Visual Comparison Of Speech (VICOS)*, a method and tool originally developed for comparing pathological and healthy speakers [1], can be used to identify such correlates. VICOS characterises speaker groups by visualising prototypical realizations of each group as well as noticeable differences between the groups. It does so *locally*, so that differences can be related to individual phonemes, which facilitates interpretability. All recordings must contain the same word sequence; thus, repetitions, insertions and deletions cannot be studied.

2. Method and Results

Using penalised [1] dynamic time warping (DTW), we establish a common time basis – relative to a 'reference' recording. We calculate loudness and spectrogram, and project these time series onto the 'timing' of the reference utterance. Local tempo variations are obtained by counting inserted and deleted frames in the alignments. Spectrogram and loudness are normalised, spectrogram and tempo are smoothed. The now fixed-length, directly corresponding time series are used to generate *prototypical realizations* (average within each group) and *within-group variability* (standard dev. within each group). The *effect size* of group affiliation is measured by Cohen's d [2] (can always be related to significance for constant groups, e. g. for 2x20 persons, $|d| = 0.8$ corresponds to $p = 0.02$, two-sided t-test).

*The research leading to these results has received funding from the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant 01IS07014B (C-AuDiT), and the German Ministry of Economics (BMWi) under grant KF2027104ED0 (AUWL).

We use a sentence from the ISLE corpus [3]: *We're planning to travel to egypt for a while or so*. Excluding reading errors, we obtained 19 German speakers (7f, 12m) and 22 Italian speakers (4f, 18m). We omit tempo and spectrogram here; in loudness, cf. Figure 1, idiosyncrasies are identifiable, for example, German speakers seem to produce the plosive /t/ more articulate (steeper slope of the mean, and positive effect size in each second half); the syllable /iː/ in Egypt, bearing both phrase and word accent, is louder in Italian (blue effect size).

3. Conclusions

VICOS is a generic system for rapidly assessing systematic differences between speakers on the basis of possibly large datasets, in an objective, interpretable and quantifiable way. We showed that it can be applied successfully for studying non-native speaker groups, too. Current work includes pitch and re-synthesis, and the usage of the projected time series as high-performance, interpretable features for automatic classification.

4. References

[1] S. Wankerl, F. Hönig, A. Batliner, J. R. Orozco-Arroyave, and E. Nöth, "Visual comparison of speaker groups," in *INTERSPEECH 2015 (Show and Tell)*, 2015, to appear.

[2] R. Coe, "It's the effect size, stupid," in *Ann. Conf. of the British Educational Research Assoc., 2002, Exeter*. [Online]. Available: http://www.leeds.ac.uk/educol/documents/00002182.htm

[3] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in *Proc. LREC*, Athens, 2000, pp. 957–964.

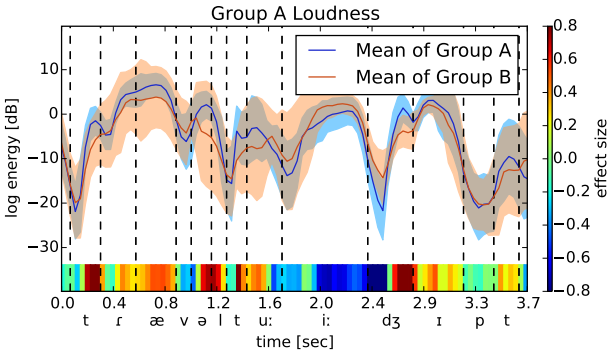

Figure 1: *Loudness: typical realizations and differences (A = German, B = Italian speakers) for the phrase "travel to egypt". Solid lines = average, semi-transparent tubes = standard deviation. Bars at bottom = effect size (yellow/red = positive ≙ higher in German; cyan/blue = negative ≙ lower in Italian).*