Visual Comparison of Speaker Groups*

Sebastian Wankerl[1], *Florian Hönig*[1], *Anton Batliner*[1,2], *Juan Rafael Orozco-Arroyave*[1,3], *Elmar Nöth*[1]

[1]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[2]Machine Intelligence & Signal Processing Group, TUM, Munich, Germany
[3]Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia

{sebastian.wankerl, florian.hoenig}@fau.de

Abstract

We describe a generic tool for visualising differences between two groups of speakers who produce a given word sequence. We do this by first time-aligning all recordings and then aggregating time-varying information within each group. By that, we can display prototypical loudness and tempo contours, and also spectrograms, together with information on variability and group effect size over time. An optional user-supplied segmentation (just needed for one of the recordings) can be used to relate local differences to individual phonemes. The system is validated with a group of speakers with Parkinson's disease and an age-matched control group. It will be provided as an open-source software package to the community.

Index Terms: paralinguistics, atypical speech, pathological speech, visualization, interpretation, acoustic features

1. Introduction

When designing an automatic system for detecting or assessing the degree of atypical speech, it is helpful to obtain a qualitative description in what way this speech deviates from other speech. This characterization can be used to build specialised acoustic features, or specialised statistical models, for improving the reliability of the system; it can also be helpful in order to explain and make credible the apparent success of given systems that rely on generic acoustic features. Characterizations of atypical speech are often available as stereotypes, which may be imprecise and unreliable; if there is pertinent literature it can be contradictory or incomplete. For a given database, one can listen through the recordings and obtain a subjective impression; it is however hard to convert it to a provable characterization. Data-driven methods can also be applied; for example, automatic feature selection can identify a small subset of sufficient features from a comprehensive, generic acoustic feature set. However, the selected features are typically complex and at the same time opaque [1]. One reason for poor interpretability is that generic acoustic features are not very specific in their scope: global features are typically more stable and easier to extract; referring to specific phonemes would give rise to the problem of sparse data, and blow up even more the typically already huge feature sets. In this paper, we describe *Visual Comparison of Speaker Groups (VICOS)*, a method and tool that helps characterising differences between speaker groups by visualising prototypical realizations of each group as well as noticeable differences between the groups. It does so *locally*, so that these differences can be related to individual phonemes, which facilitates interpretability. The method is generic and can be applied to all kinds of speech, with the restriction that all recordings must contain the same word sequence. Thus, repetitions, insertions and deletions cannot be be studied; pauses inserted by just a subset of the speakers may complicate interpretation.

*The research leading to these results has received funding from the Hessen Agentur, grant numbers 397/13-36 (ASSIST 1) and 463/15-05 (ASSIST 2). Juan Rafael Orozco-Arroyave is under grants of COLCIENCIAS through the program "Convocatoria Nº 528, generación del bicentenario 2011".

2. Method

We establish a mapping that tells us when one speaker was pronouncing the same bit of speech as another speaker. To create this *alignment*, we apply dynamic time warping (DTW). It has the advantage of being suitable for any language and speech atypicality. We use the standard short-time representation Mel Frequency Cepstral Coefficients (MFCC). Channel and global speaker characteristics are removed by mean subtraction. To add context information, we augment each cepstral coefficient with its derivative (slope of best fit through 5 frames). For some kinds of atypical speech, these derivatives can have a negative impact; therefore, we also offer a set-up without them. To improve the quality of the alignments, we use a penalty for each inserted or deleted frame [2, 3]. To balance penalty and distance in the space of MFCC + derivatives, we normalize each coefficient and derivative to standard deviation 1. Thus, we found a penalty of 1 per inserted/deleted frame suitable for good alignments; for the set-up without derivatives, 5 turned out to be appropriate. Apparently, the derivatives have a similar regularization effect as the insertion/deletion penalty, and without the derivatives, the insertion/deletion regularization needs to be reinforced. We align all recordings to a single 'reference' recording selected by the user. These alignments now constitute the common time basis; the user can provide a segmentation of the reference utterance to locate individual phonemes. We now calculate time-varying quantities of interest from the recordings. Currently, we provide loudness and spectrogram. Using the alignments, the time series are then projected onto the 'timing' of the reference utterance. We also compute local tempo variations relative to the reference utterance by counting inserted and deleted frames in the alignments. After projection, spectrogram and loudness are normalised and partly smoothed (spectrogram: Gaussian filter with σ=5 frequency bins $\approx$ 108 Hz). The logarithm is only taken afterwards to concentrate on speech rather than silence in the normalization, and to obtain an 'upper envelope' effect in the spectrogram smoothing like in MFCC computation [4, Fig. 5 (ii)]. The local tempo is also smoothed, with σ=2.5 frames (25 msec).

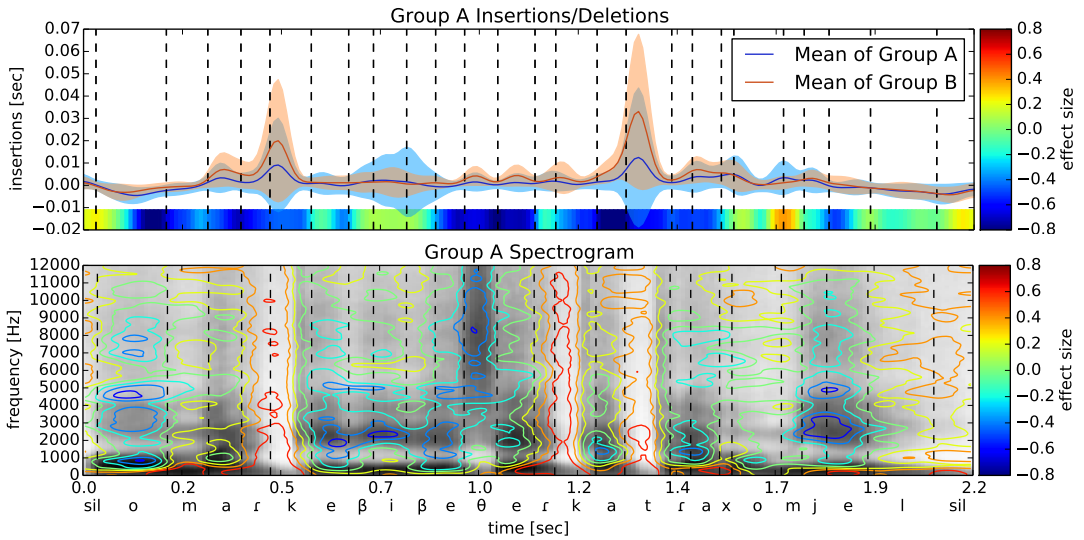

Figure 1: *Prototypical realizations and differences for group A (Parkinson's disease) versus group B (controls). For inserted/deleted frames (upper graph), the average of each group is plotted as a solid line, along with a semi-transparent tube for the standard deviation; effect size is indicated by bars at the bottom (red = positive $\hat{=}$ higher in PD; blue = negative $\hat{=}$ lower in PD). For the spectrogram, the average is displayed just for A (in grey levels); effect size is shown by contour lines (again, red = higher in PD; blue = lower in PD).*

We now obtained fixed-length, directly corresponding time-varying signals for all recordings. We generate *prototypical* realizations from these by averaging all signals within each group. Within-group variability is given by the signal's standard deviation across speakers. The effect of group affiliation can be assessed by comparing the difference of the group means μ_A, μ_B with the standard deviations σ_A, σ_B. We chose Cohen's d as *effect size* measure, as the more relevant counterpart of significance [5], as it does not depend on the group size: $d = (\mu_A - \mu_B)/\sqrt{(\sigma_A^2 + \sigma_B^2)/2}$; $|d| = 0.8 \approx$ strong effect; $0.5 \approx$ moderate; $0.2 \approx$ weak. Note that for constant groups, effect size can always be related to significance, e. g. for 40 persons in each group, $|d|$=0.5 corresponds to p=0.03 (two-sided t-test). The effect size for the spectrogram is smoothed across time with a Gaussian filter ($\sigma = 2.5$ frames $\hat{=}25$ msec).

3. Experiments and Results

We test VICOS with Spanish speech from 50 speakers (25 f, 25 m) suffering from Parkinson's disease (PD) and 50 age-matched controls (25 f, 25 m) [6]. We use the sentence *Omar, que vive cerca, trajo miel* (Omar, who lives nearby, brought honey). Removing files with reading errors left us with 42 PD speakers (21f, 21m) and 49 healthy controls (25f, 24m). A number of effects characterising PD [7] can be observed in Figure 1 (loudness is omitted for space reasons; it is comparable to the visualization of tempo): (1) rushed speech (more blue in the bottom of insertions/deletions = fewer inserted frames = increased tempo); (2) slowing down at phonemes difficult for PD speakers (green and orange during /β i β e/, /r k/, and /om/); (3) less control of fricative (light blue contour line in spectrogram 7000–10000 Hz during /θ/); (4) impaired control of velum: higher energy across the spectrum before closures (red/orange contour lines before each /k/); (5) less control on tongue: less energy of /t/ and /l/ across spectrum (orange contour lines during /t/ and /l/); (6) more nasality: generally in vowels, more energy in the frequency band of nasal formants (red/orange contour lines below 400 Hz, blue contour lines above).

4. Conclusions

VICOS can rapidly uncover systematic differences between speaker groups—in an objective, quantifiable, and interpretable manner. Application to PD speech reproduced several characterizations known from literature. Some caveats have to be made: due to the large number of parameters analysed, interpretation should be careful, especially when sample size is small. Moreover, violations of the normality assumption might distort results. Any speaker group comparison, and other fields such as phonetics should benefit from VICOS. We are currently integrating pitch, and resynthesis of prototypical realizations. We also plan to use the projected data (time-normalised tempo contour, loudness contour, and spectrogram) as comprehensive but still interpretable features for classification.

5. References

[1] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Acoustic-Prosodic Characteristics of Sleepy Speech — between Performance and Interpretation," in *Proc. of Speech Prosody 2014*.

[2] K. Roberts, P. Lawrence, A. Eisen, and M. Hoirch, "Enhancement and dynamic time warping of somatosensory evoked potential components applied to patients with multiple sclerosis," *Biomedical Engineering, IEEE Transactions on*, no. 6, pp. 397–405, 1987.

[3] H. Sun, J. C. Lui, and D. K. Yau, "Distributed mechanism in detecting and defending against the low-rate TCP attack," *Computer Networks*, vol. 50, no. 13, pp. 2312–2330, 2006.

[4] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising perceptual linear prediction (PLP)," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 2997–3000.

[5] 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

[6] J. Orozco-Arroyave, J. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease," in *LREC*, 2014, pp. 342–347.

[7] U. Jürgens, "Neural pathways underlying vocal control," *Neuroscience & Biobehavioral Rev.*, vol. 26, no. 2, pp. 235–258, 2002.