Audio feature space extraction for dysarthria assessment

Acoustic analysis of dysarthric speech is typically conducted either manually, with the researcher selecting the signal stretches to be used for the measurement of some parameter (e.g. middle portion of vowel for jitter and shimmer analysis), or automatically with the extraction of large acoustic feature spaces and their consequent input into a machine learning system. A recent example of this type of analysis is the 2015 Interspeech Computational Paralinguistics Challenge on Parkinson's Disease (Schuller et al. 2015) where acoustic features were automatically extracted with *openSMILE* (Eyben et al. 2013) and continuosly fed into *Weka* (Witten & Frank 2005). Both approaches have their obvious drawbacks: manual acoustic analyses are work-intensive and error-prone, whereas automatic machine learning requires large amounts of data and does not deliver explicit knowledge on the most relevant combinations of parameters.

The present study is an attempt to combine both methodologies in order to minimize manual interaction with the data and at the same time maximize the amount of explicit knowledge gained. Recordings from the eight speakers of the Nemours database (Menendez-Pidal et al. 1996) were analyzed with *openSMILE*, using the 1582 low-level and derived acoustic features originally applied to the Interspeech 2010 Paralinguistic Challenge (Schuller et al. 2013), and the resulting output searched for correlations with the reported overall Frenchay assessment scores, calculated as the mean of the eight sub-scores. Out of the one hundred most highly correlated features various sets of five to twenty features were created and tested for their predictive power with linear regression models. Encouraging results were obtained for models with as few as ten features predicting correctly the rank order of the speakers' overall assessment scores. Almost all of these models included several f_0-related parameters in addition to parameters calculated from mel-frequency cepstral coefficients and they all made use of derived statistical *openSMILE* parameters like *upleveltime75* or *percentile1.0*.

But there were also problems: there was considerable variation in the correlation of individual recordings' analyses with the assessment scores and there seemed to be an effect of recording duration on predictive power: the worst predictions tended to come from very short and very long recordings. Obviously, more data is needed to address these problems and to collect more evidence for the best feature combinations. Also separate prediction of scores for individual physiological subsections of the Frenchay assessment has not yet been tried.

References

Eyben, F., Weninger, F., Gross, F., Schuller, B. 2013: "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", In *Proc. ACM Multimedia (MM)*, Barcelona, Spain, ACM, pp. 835-838.

Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzjo,J. E., Bunnell, H. 1996: "The Nemours Database of Dysarthric Speech". *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia PA, USA, pp. 1962–1965.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S. 2013: "Paralinguistics in Speech and Language – State-of-the-Art and the Challenge," *Computer Speech and Language*, Special Issue on Paralinguistics in Naturalistic Speech and Language, vol. 27, no. 1, pp. 4–39.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y., Weninger, F. 2015: "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition", *INTERSPEECH-2015*, pp. 478–482.

Witten, I., Frank, E. 2005: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Ed., San Francisco: Morgan Kaufmann.