Spectral Tilt Used for Automatic Detection of Prominence: A Comparison Between Electroglottography and Audio

Laura Enflo

(Term Paper for Machine Learning, a GSLT course 2009)

Introduction

Words of particular significance in sentences are often spoken with emphasis, or prominence. Hence, knowing how to automatically find these particular terms would be a desirable objective. It has previously been suggested that word and syllable duration and fundamental frequencies (F0) are important cues for prominence detection [for Swedish, e.g. 1 & 2]. Other investigated parameters are audio spectral tilt and loudness, but both have been shown to contribute less to the detectability than duration and F0 [1 & 3].

Electroglottography (EGG) is a widely used technique for the assessment of vocal-fold contact during phonation [e.g. 4]. An electroglottograph is provided with two electrode plates, which are placed on each side of the larynx. The idea with EGG is to send an electrical signal from one electrode to the other and record the amplitude of this signal. When the vocal folds are closed, the signal can pass, such that the amplitude is high, and when the glottis is open, the amplitude is low. Several studies have shown that EGG is a more robust technique than audio for F0 determination [e.g. 5], partly due to its higher correlate to vocal source characteristics.

Spectral slope, or spectral tilt, is often defined as the slope of least squares linear fit to the log power spectrum. A lower spectral tilt corresponds to a louder voice and when loudness decreases, the spectral slope has been confirmed to increase [e.g. 6 & 7].

This course project aims to compare the importance of the two syllable-level features EGG versus audio spectral tilt in automatic prominence detection.

Data and Annotation

In this study, 200 out of 5000 sentences were chosen from a dataset containing recorded speech read by a male professional Swedish actor. The corpus includes both audio and corresponding EGG signals. The selected sentences were then annotated according to the level of prominence perceived by four speech experts, who marked each word as prominent, not prominent or maybe prominent. The average answer of the four subjects was rounded into three levels: 0 (No prominence) when x<0.5, 1 (Maybe prominent) when $0.5 \leq x \leq 1.5$, and 2 (Prominent) when $x>1.5$ [1]. Out of the 200 sentences, 150 were randomly chosen for training and 50 for testing.

Experiment

The vowels A, E, I, O, U and Y were identified in and picked from the dataset with prominence-labeled words, however under the condition that the syllable had a length of at least 600 data points, thus eliminating less reliable data. One spectral tilt value for each of the audio and EGG signals was calculated by means of the 600 data points in the middle of each syllable, in a straight-forward fashion as mentioned in the introduction. A memory-based learner was then used, namely the Tilburg Memory-Based Learner (TiMBL), which is a decision-tree-based implementation of k-nearest neighbor classification [1 & 8]. The k-value was chosen to be 20, which was the best value according to the similar study of [1].

Results and Discussion

The test data gave a baseline word accuracy of 40.8%, which is considerably lower than that of a previous study on the same dataset which contained more data points [1]. In the older investigation, the syllable-level audio spectral tilt accuracy was 42.8% as compared to the baseline word accuracy of 52.9%. However, since this study was aimed for a comparison between parameters, the lower value of the baseline word accuracy in this project was not considered overly problematic. On the other hand, the comparison showed a rather surprising equality between the two investigated parameters. Although the EGG spectral tilt was at first considered to be a more promising parameter than the audio spectral tilt, the results show small differences between the two. The overall accuracy was 27% for the EGG spectral tilt and 31% for the audio spectral tilt parameter. A combination of the two parameters gave a confusion matrix as seen in Table 1, which shows that the detection of no prominence can be considered as especially problematic.

Table 1. Confusion matrix for the two syllable-level parameters EGG and audio spectral tilt used simultaneously with TiMBL.

Predicted Class	No Prominence	Maybe Prominence	Prominence
True Class			
No Prominence	7.56	42.86	49.58
Maybe Prominence	4.21	12.63	83.16
Prominence	8.97	12.82	78.21

As previously mentioned, the spectral tilt has been shown to be closely connected to the loudness. This could explain the poor results, since also loudness has a limited impact on prominence as seen in [1]. In both the present and the previous study, the spectral tilt parameter accuracies have been lower than the baseline word accuracy. This supports the conclusion that spectral tilt parameters, both the EGG and the audio versions, do not play key roles in automatic prominence detection.

References

[1] Al Moubayed, S., Ananthakrishnan, G. & Enflo, L. (forthcoming) Automatic Prominence Classification in Swedish.

[2] Fant, G., Kruckenberg, A. & Liljencrants, J. (2000) Acoustic-phonetic analysis of prominence in Swedish. In: Botinis A (ed.), Intonation: Analysis, modeling and technology, Dordrecht: Kluwer Academic Publishers, pp. 55-86.

[3] Campbell, N. & Beckman, M. (1997) Stress, prominence and spectral tilt. In A. Botinis, G. Kouroupetroglou and G. Carayannis (eds), Proceedings of the ESCA Workshop on Intonation: Theory, Models and Implications, Athens, Greece, September 18-20, pp. 67−70. Athens: ESCA and University of Athens.

[4] Orlikoff, R. F. (1998) The uses and abuses of electroglottography, Phonoscope 1, pp. 37–53.

[5] Vieira, M. N., McInnes, F. R. & Jack, M. A. (1996) Robust F0 and jitter estimation in pathological voices. In ICSLP-1996, 745-748.

[6] Hanson, H.M (1997) Glottal characteristics of female speakers: Acoustic correlates. Journal of the Acoustical Society of America 101(1), 466-481.

[7] Fant, G. & Lin, Q. (1988) Frequency domain interpretation and derivation of glottal flow parameters. Speech Transmission Laboratory Quarterly Progress Scientific Report 1988(2-3), 1-21.

[8] Daelemans, W., Zavrel, J., van der Sloot, K. & van den Bosch, A. (2007) Timbl: Tilburg memory-based learner, version 6.2.0, http://ilk.uvt.nl/software/