Creaky voice (or vocal fry) is a voice quality frequently produced in many languages. In order to enhance the naturalness of speech synthesis, a proper use of creaky voice should be included. The goal of this paper is two-fold:

1) Which contextual factors are the most relevant in predicting creaky usage
2) To what extent contextual factors can be useful for the prediction of creaky voice from contextual factors

The aim here is to investigate:

1) Which contextual factors are the most relevant in predicting creaky usage
2) To what extent contextual factors can be useful for the prediction of creaky voice from contextual factors

For US English, the standard complete list of 53 contextual factors in the HTS implementation [2, 3] are used, relating to phoneme, syllable, word, phrase, utterance type and position. The predictability power of each contextual factor was assessed based on its mutual information (MI) with the creaky use decisions. Only 13 contextual factors are found to have interesting normalised MI values higher than 15%. The contextual factors are closely related with creaky use at the end of a sentence or a word group.

Examples of contextual factors:
- Phoneme:
  - preceding, current, succeeding phonemes
  - position of current phoneme in current syllable
- Syllable:
  - no. of phonemes at (preceding, current, succeeding) syllable
  - onset of (preceding, current, succeeding) syllable
  - stress of (preceding, current, succeeding) syllable
  - position of current syllable in word
- Word:
  - part of speech of (preceding, current, succeeding) word
  - number of syllables in (preceding, current, succeeding) word
  - number of words from previous to next content word
  - Place:
    - number of syllables in (preceding, current, succeeding) phrase
  - Utterance:
    - number of syllables in current utterance

The analysis shows that a few contextual factors, related to speech production preceding a silence or a pause, are of particular interest for prediction.

In the second experiment, four methods are proposed to predict the use of creaky voice based on HMMs. It is shown that modelling the posterior probability given by the detection algorithm leads to the best results across all metrics. This technique achieves performance scores comparable to the determination rates obtained by the detection method on which it is trained.

Acknowledgements This research was supported by FNRS, the Science Foundation Ireland (FastNet), the Irish Department of Arts, Heritage and the Gaeltacht (ABAIR project), the EU’s FP7 project Simple4All, Academy of Finland, and Aalto University’s MIDE UI-ART project.

Contact: thomas.drugman@umons.ac.be, kanejo@tcd.ie, tuomo.rawilo@aalto.fi

References

4 Creaky voice prediction methods based on HMMs

Four different creaky voice prediction methods are experimented with. The creaky voice related features are trained along with the conventional HMM-based synthesis features, F0 and spectrum.

I. PredictedFeat The two features given by the creaky detection algorithm are trained in two separate streams, after which the prediction is drawn from the decision tree used in detection method.

II. Binary The binary decision output of the creaky detection algorithm is trained in continuous stream. The final decision is made by thresholding the trained probability with a pre-specified value.

III. Binary MSD The binary decision output of the creaky detection algorithm is trained in multi-space probability distribution stream, aligned with F0. The final decision is the stream output.

IV. PosteriorProb The posterior probability given by the creaky voice detection algorithm is trained in a continuous stream. The final decision is made by thresholding the trained probability.