PREDICTION OF CREAKY VOICE FROM CONTEXTUAL FACTORS

UMONS Université de Mons
Thomas Drugman
TCTS Lab
University of Mons, Belgium

TRINITY COLLEGE DUBLIN
John Kane
Phonetics and Speech Laboratory
Trinity College Dublin, Ireland

Tuomo Raitio
Department of Signal Processing and Acoustics
Aalto University, Espoo, Finland

Summary:

- **Creaky voice is voice quality frequently produced in many languages**
- **The analysis shows that a few contextual factors, related to speech production preceding a silence or a pause, are of particular interest for prediction**
- **Four prediction methods based on training and generating a creaky probability stream with HMMs are compared on US English and Finnish speakers**
- **The best prediction technique performs comparable to the creaky detection algorithm on which HMMs were trained**

1 Introduction

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. Analyse how informative contextual factors are in term of predicting creaky voice

2. Investigate various creaky voice prediction schemes based on HMMs

2 Creaky voice detection

An automatic creaky voice detection technique, described in [1] is used. The algorithm involves the use of two acoustic features which characterise two different aspects of the creaky excitation:

– H2–H1 of a resonator output

– Residual peak prominence

These features are used as part of a decision tree classifier for binary creaky decision.

3 Analysis of contextual features related to creaky voice

The aim here is to investigate:

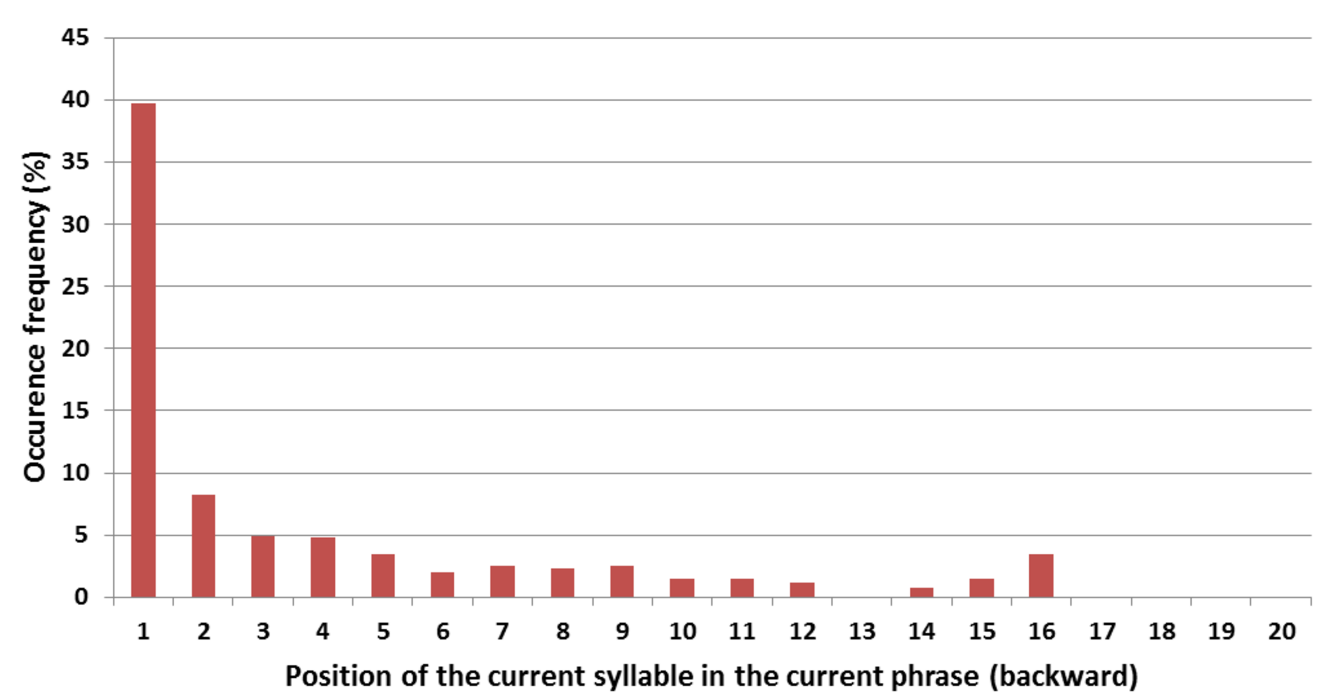
1) Which contextual factors are the most relevant in predicting creaky usage

2) To what extent contextual factors can be useful for the prediction

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} phoneme
  - position of current phoneme in current syllable
- Syllable
  - no. of phonemes at {preceding,current,succeeding} syllable
  - accent 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
  - position of current word in current phrase
  - number of words {from previous, to next} content word
- Phrase:
  - number of syllables in {preceding, current, succeeding} phrase
- Utterance:
  - number of syllables in current utterance



4 Creaky voice prediction methods based on HMMs

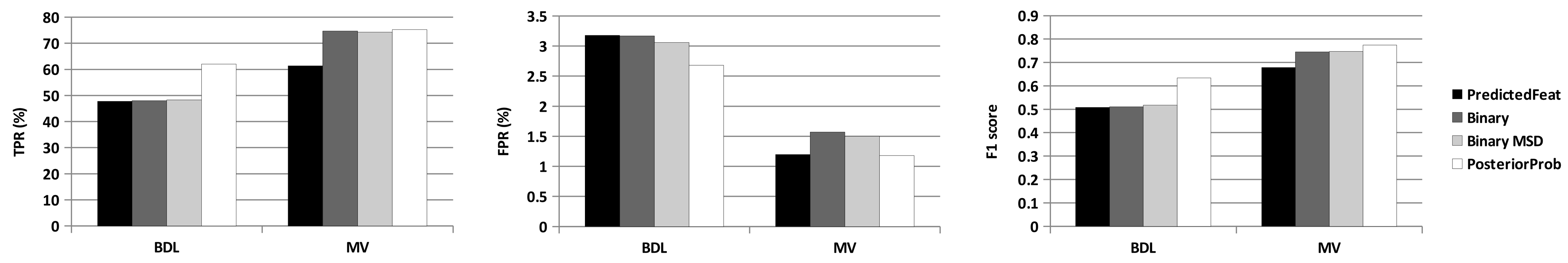
Four different creaky voice prediction methods are experimented with. The creaky voice related features are trained along with the conventional HMM-based synthesis features, F_0 and spectrum:

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 F_0. 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.



5 Results

The methods are tested on US English (BDL) and Finnish (MV) databases. Frame-level metrics, true positive rate (TPR or recall), false positive rate (FPR), and F1 score are evaluated. Across both speakers and all metrics, the PosteriorProb method gives the best performance.

Database	Method	Misses	FAs	Hits
BDL	PredictedFeat	66	24	98
	Binary	68	19	96
	Binary MSD	68	**17**	96
	PosteriorProb	**40**	37	**124**
MV	PredictedFeat	54	29	109
	Binary	46	25	117
	Binary MSD	47	**24**	116
	PosteriorProb	**28**	39	**135**

6 Conclusions

Firstly, it has been investigated how how contextual information is related to the use of creaky voice. Contextual factors linked to speech production preceding a silence or a pause appears to be highly relevant, leading to normalised mutual information values up to 32%. This confirms that vocal fry has a syntactic role by making a better delimitation of groups of words and by making phrase segmentation easier.

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.

References

[1] Kane, J., Drugman, T., Gobl, C., "Improved automatic detection of creak", Computer Speech and Language, 27(4):1028–1047, 2013.

[2] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. and Tokuda, K., "The HMM-based speech synthesis system (HTS) version 2.0", in Sixth ISCA Workshop on Speech Synthesis, Aug. 2007, pp. 294–299.

[3] [Online] "HMM-based speech synthesis system", *http://hts.sp.nitech.ac.jp*.

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.raitio@aalto.fi