Detecting Non-modal Phonation in Telephone Speech

Tae-Jin Yoon[1], *Jennifer Cole*[2], *& Mark Hasegawa-Johnson*[3]

Department of Linguistics[1,2]; Department of Electrical and Computer Engineering[3]
University of Victoria, Canada[1]; University of Illinois at Urbana-Champaign, U.S.A.[2,3]
tyoon@uvic.ca; {jscole; jhasegaw}@uiuc.edu

Abstract

Non-modal phonation conveys both linguistic and paralinguistic information, and is distinguished by acoustic source and filter features. Detecting non-modal phonation in speech requires reliable F0 analysis, a problem for telephone-band speech, where F0 analysis frequently fails. We demonstrate an approach to the detection of creaky phonation in telephone speech based on robust F0 and spectral analysis. Our F0 analysis relies on an autocorrelation algorithm applied to the intensity-boosted and inverse-filtered speech signal and succeeds in regions of non-modal phonation where the non-filtered F0 analysis typically fails. In addition to the extracted F0 values, spectral amplitude is measured at the first two harmonics (H1, H2) and the first three formants (A1, A2, A3). Visual and spectral inspection of the detected creaky phonation confirms the findings reported from laboratory setting. Statistical analysis using one-way ANOVA and classification using Support Vector Machine (SVM) reveals promising results which lead to further improvement for automatic detection of non-modal phonation in telephone speech.

1. Introduction

Through modulation in source and filter characteristics, speech conveys both linguistic and paralinguistic information. Fundamental frequency (F_0) and harmonic structure are important factors in encoding lexical contrast and allophonic variation related to laryngeal features [1][2][3]. They also play an important role in the expression of prosodic features of stress and intonation [4][5][6]. In addition, shifts in F_0 and voice quality can signal emotional state or affect, as for example the expression of boredom signalled by creaky voice, or intimacy signalled by breathy voice [7].

It has been widely noted that there is a relationship between F_0 and voice quality. For example, Maddieson and Hess [3] observe significantly higher F_0 for tense vowels in languages that distinguish three phonation types (Jingpho, Lahu and Yi) and Ohala [2] observes that Hindi breathy voiced stops are distinguished by a lowered F_0 at voice onset following release. However, F_0 is not always a strong indicator of voice quality, as shown by studies of English that fail to show a strong correlation between any glottal parameters and F_0 [8][5]. On the other hand, information obtained from spectral structure has been shown to be more reliable for the discrimination of non-modal from modal phonation. For example, Gordon and Ladefoged [1] describe the characteristics of creaky phonation as having non-periodic glottal pulses, lower power, lower spectral slope, and low F_0. They report that spectral slope is the most important feature for discrimination between different phonation types. Ní Chasadie and Gobl [9] also characterize creaky phonation as having extremely low F_0 and irregular glottal pulses. They state that significant spectral cues to creaky phonation are i) a very dominant A1 (i.e., amplitude of the strongest harmonic of the first formant) relative to H1 (i.e., amplitude of the first harmonic), and ii) H2 (i.e., Amplitude of the second harmonic) higher than H1.

The studies cited above establish the importance of spectral structure, and in particular the relative amplitude of the lower harmonics, for voice quality identification. Extraction of harmonic structure, in turn, requires reliable extraction of F0. For low fidelity recordings, such as those characteristic of telephone speech, reliable F0 analysis is a major challenge. An added problem in the analysis of spectral structure is the effect of pitch variation; spectral measurements are sensitive to differences in pitch that can amplify or attenuate the amplitude of some harmonics [5]. Previous studies find that reliable voice quality measures require high fidelity recording, ideally supplemented by EGG data to support F0 analysis, and a labor-intensive process of manual, interactive analysis (e.g., [5] [9]). This state of affairs calls to question the viability of voice quality analysis for large speech corpora, especially corpora consisting of low quality recorded speech, such as telephone speech. We address this challenge in the present study, which presents a method for the extraction of F0 and harmonic structure that proves effective for telephone speech, and results from a classification experiment that demonstrate the usefulness of these acoustic features for voice quality identification. These issues are discussed as follows. Sections 2 introduces methods for extracting F0 and spectral features from the Switchboard corpus of telephone conversation speech. Section 3 presents ANOVA results that establish a relationship between F0 and spectral features and perceived voice quality in the Switchboard corpus, and results from classification experiments using Support Vector Machines to identify creaky and modal voice quality, based on perceptual labels. Section 4 concludes the paper with discussion of strategies for improving the differentiation of non-modal phonation from modal phonation in telephone speech.

2. Method

2.1. Corpus

Switchboard is a corpus of spontaneous telephone conversations between strangers [11]. The corpus is designed mainly to be used in developing robust speaker-independent Automatic Speech Recognition (ASR). In general, the quality of the recorded speech, which is sampled at 8kHz, is much inferior to speech samples recorded in the phonetics laboratory. As noted by Taylor [12], pitch tracking algorithms known to be reliable for laboratory-recorded speech (e.g., [4]) often fail to extract an F_0 during regions perceived as voiced from Switchboard corpus. Our analysis is based on the Switchboard files

in the WS97 subset for which there are hand-corrected word- and phone-aligned transcriptions. Through the course of a transcription project (not reported here), about 200 WS97 files were identified as containing regions of non-modal, creaky phonation. Phonation quality was assessed on the basis of visual inspection of the spectrogram and waveform, and listening. Out of these 200 files, we selected 160 intervals of modal phonation and 140 intervals of creaky phonation for acoustic feature analysis and classification experiments.

2.2. Feature extraction: F0 and spectral cues

Extraction of F0 and spectral features is done using Praat [14] following the algorithm diagrammed in Figure (1). First, uniform intensity normalization based on 85 dB is applied to each file. Intensity normalization is necessary because the level of intensity in Switchboard sometimes falls below the threshold necessary for F0 analysis. Following intensity normalization, inverse filtering is applied and F0 analysis is calculated on the intensity-normalized, inverse-filtered signal, using the autocorrelation method developed by [13] in a window that is dynamically sized to contain at least four glottal pulses based on the minimum F0 obtained from unfiltered AC analysis. Harmonic structure is determined through spectral analysis using FFT and long term average spectrum (Ltas) applied to the intensity-normalized, inverse filtered signal. H1 and H2 are calculated through integer multiplication of the F0 value obtained from the autocorrelation analysis. A1 through A3 measures are obtained on the basis of the intensity normalized signal without inverse filtering. After postprocessing formant values using the Praat backtracking features, spectral analysis using FFT and Ltas is performed to obtain the formant amplitude measures A1, A2 and A3, and spectral slope. The F0 and spectral features obtained from these two analyses of each windowed frame are combined in the calculation of H1-H2, H1-A1, H1-A2, and H1-A3, and these measures along with F0 and spectral slope are the basis for our classification experiments. Intensity measures are not directly evaluated for this experiment. The measures are extracted from windowed signals at 50 ms. intervals.

The extracted F0 and spectral features are evaluated in relation to the perceptually-based labels of creaky vs. modal phonation for a total of 300 sound intervals in the 200 files we selected from in the WS97 subset. While this study is focused on the acoustic correlates of perceived creaky voice in our data, we recognize that other types of non-modal phonation such as tenseness or pressed voice may also occur in the corpus without explicit labeling. Figure (2) presents an example waveform of the utterance *yeah that* that is identified as exhibiting creaky or glottalized voice quality in the time-indexed interval from 0.665 to 0.765 seconds. F0 and spectral features are taken from two to three windowed samples (depending on the length of the interval) within the interval labeled creaky and logged into the creaky phonation database, and features from another two to three windowed samples are taken from modal phonation regions of the same file and logged into the database for modal phonation. To balance the distribution of creaky and modal phonation samples across the corpus, features from no more than three feature windowed samples are extracted from the same file for creaky and modal voice qualities. Figure (3) shows that changes in F0 (top), H1-H2 (mid), and H1-A1 (bottom) values occur at the juncture between the labeled creaky and modal voice intervals in Figure (2).

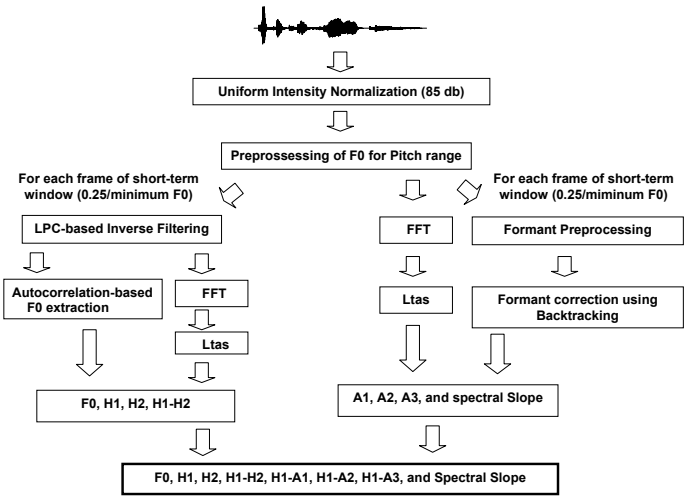

Figure 1: *Diagram of feature extraction algorithm*

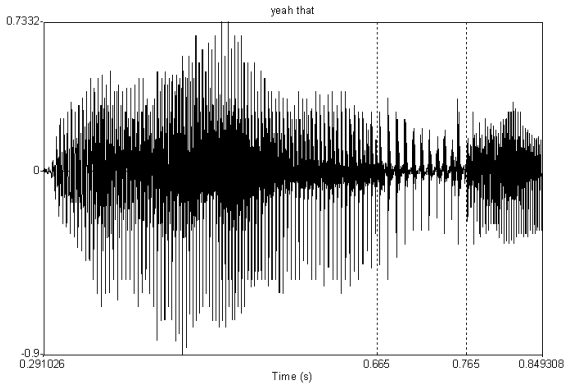

Figure 2: *Waveform of the utterance 'yeah that' with perceived creaky phonation or glottalization in the time interval from 0.665 to 0.765 seconds*

3. Analysis and classification

3.1. One-way ANOVA

Figure (3) shows that the regions perceived as instances of creaky voice are associated with low F_0 and relatively high H2 compared to H1. Results from separate one-way ANOVAs for each acoustic feature with the independent factor Phonation Type (creaky, modal) are given in Table (1), with degree of freedom of 1 for between groups and 298 for within groups. These results show that a significant effect of Phonation Type for every feature $H2$. These results based on perceptual labeling of prototypical creaky and modal exemplars are in line with findings from previous acoustic studies (e.g.,[9]).

One possible reason for the absence of a Phonation Type effect on H2 is that while the variation of H1 is large, the variation of H2 is relatively small, as illustrated in Figures 4 and 5, showing long term average spectra (Ltas) from two different speakers for modal voice (Figure 4) and creaky voice (Figure

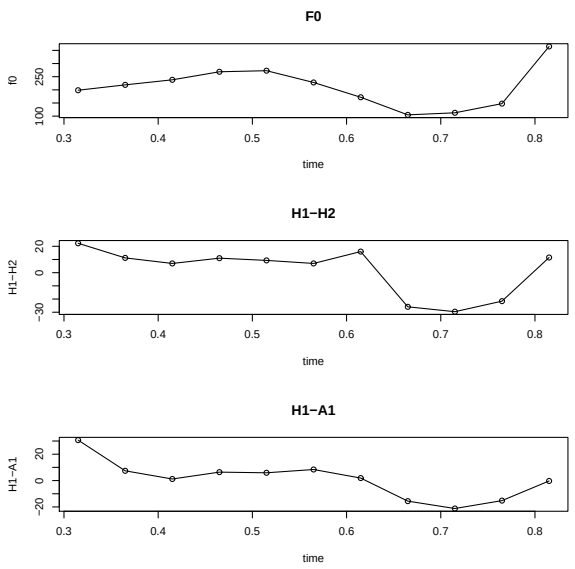

Figure 3: *Plots of F0 (above), H1-H2 (mid), and H1-A2 (below) for the utterance 'yeah that' obtained from the algorithm diagrammed in Figure (1)*

Table 1: *ANOVA Results*

	F value	Sig
F0	100.846	< 0.001
H1	210.000	< 0.001
H2	0.395	< 0.53
H1−H2	206.599	< 0.001
Slope	11.611	< 0.005
H1−A1	67.887	< 0.001
H1−A2	102.798	< 0.001
H1−A3	67.284	< 0.001

5) regions of the same word *table*. Even though the two voice qualities differ in H1-H2 values, H2 is relatively similar (around 25 to 30 dB) in both Figure (4) and Figure (5).

3.2. Classification using Support Vector Machine(SVM)

Classification experiments were conducted using the Support vector machine (SVM) classification learning algorithm using the radial basis function (RBF) availabe in libsvm [15]. Sound files from the Switchboard subset described above were classified as either creaky or modal based on the acoustic features of F0 and harmonic structure using the best parameters returned by parameter search tool in [15] (i.e., $C = 8.0$, and $\gamma = 0.125$). Classification of phonation type (creaky vs. modal) results in 75% accuracy. The result shows an improvement from the baseline of 53% (160/300).

4. Discussion and conclusion

When F0 is reliably extracted, it is possible to detect creaky phonation on the basis of harmonic structure, even on telephone speech. The analysis of creaky phonation from telephone speech is in line with the previous results obtained from laboratory speech (e.g., [9]). This promising result suggests

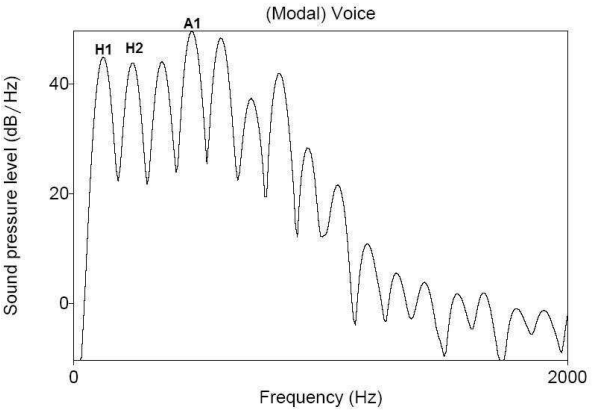

Figure 4: *H1−H2 for a modal voice sample*

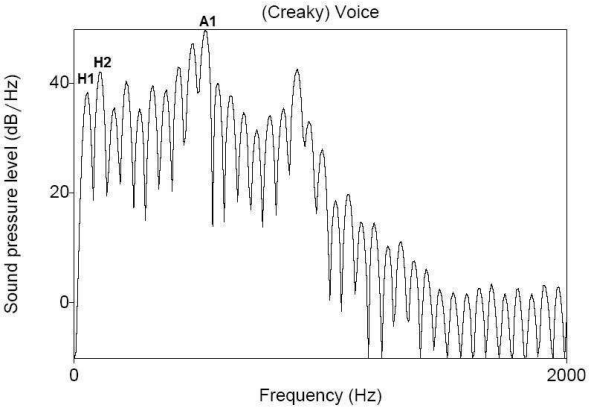

Figure 5: *H1−H2 for a creaky voice sample*

the viability of automatic detection of non-modal phonation from unlabeled speech data. However, creaky phonation is just one of many types of voice quality found in natural conversational speech. For example, tense phonation (also called pressed phonation) and creaky phonation share considerable similarity in spectral features. Ní Chasaide and Gobl[9] state that if the pulse-to-pulse variability is ignored, the distinction between tense and creaky phonation is difficult to make. Figure (6) and Figure (7) present scatterplots showing the relationship between F0 and H1-H2, and between H1 and H2, respectively. The values are taken from vocalic regions in all 200 files used for the acoustic and classification experiments discussed above. Figure (6) demonstrates a nonlinear relationship between F0 and H1-H2, which confirms earlier studies [8][5][7] in finding that F0 alone is not a good predictor of voice quality. Figure (7) reveals two clouds in the distribution of H1 and H2, and although the data in this analysis have not yet been (but soon will be) coded for phonation type, the distribution may be indicative of a split between modal and non-modal phonation. Based on descriptions by Ní Chasaide and Gobl [9], we speculate that the cloud for nonmodal phonation may in fact comprise two types of non-modal phonation: creaky and tense phonation.

The features used in this paper are primarily those extracted from the frequency domain. We are currently improving our algorithm for detecting creaky phonation by incorporating features taken from the time domain in order to discriminate creaky

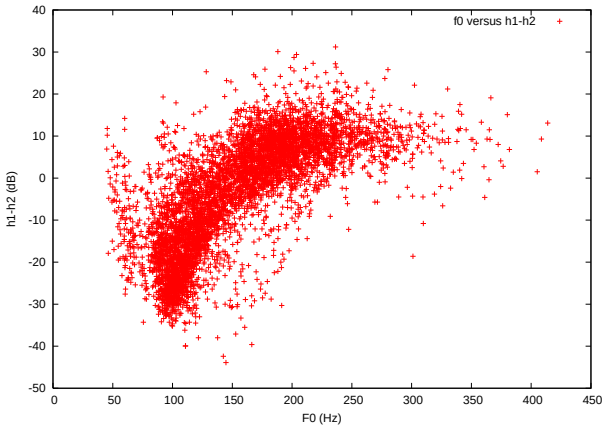

Figure 6: *F0 versus H1−H2*

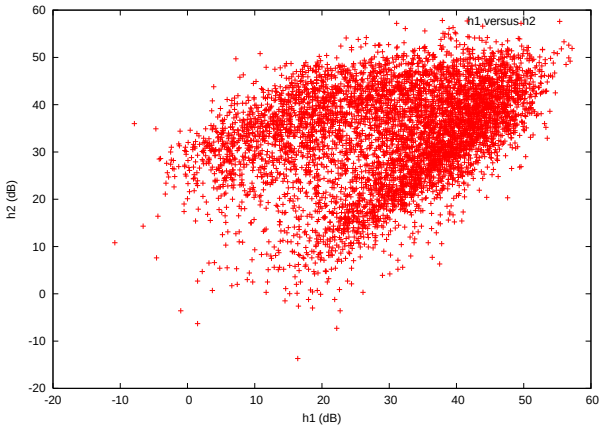

Figure 7: *H1 versus H2*

phonation from tense phonation. We are also experimenting with clustering algorithms, using the features extracted in figure (7). Results demonstrated in this paper suggest that we can use the features described here to automatically learn, in an unsupervised or lightly-supervised fashion, clusters corresponding respectively to modal and creaky phonation, and that the class boundaries learned in this way may be used for voice quality analysis in automatic speech recognition. Our goal is to develop tools for the automatic detection of acoustic features that signal prosodic events (accent and phrasal juncture) and regions of disfluency. Voice quality, and in particular glottalization, is known to play an important role in signalling these events. Thus, detection of non-modal phonation provides a basis for recognition of prosody and disfluency.

5. Acknowledgements

This work is supported by NSF award number IIS-0414117. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

6. References

[1] Gordon, M. and Ladefoged, P. "Phonation types: a cross linguistic overview". *Journal of Phonetics* 29, pp. 383-406, 2001.

[2] Ohala, J. "The Physiology of tone". *Southern California Occasional Papers in Linguistics* 1, pp. 1-14, 1973.

[3] Maddieson, I and Hess, S.A. "The effects of F0 of the linguistic use of phonation type". *UCLA Working Papers in Phonetics*, Department of Linguistics, University of California, Los Angeles, 67, pp. 112-118. 1987.

[4] Bagshaw, P.C. "Automatic prosodic analysis for computer aided pronunciation teaching" Ph.D. dissertation, University of Edinburgh, 1994.

[5] Epstein, M.A. "Voice Quality and Prosody in English" Ph.D. dissertation, UCLA, 2002.

[6] Redi, L. and Shattuck-Hufnagel, S. "Variation in the rate of glottalization in normal speakers" *Journal of Phonetics*, 29, pp. 407-427, 2001.

[7] Ní Chasaide, A. and C. Gobl, Voice quality and F0 in Prosody: Towards a Holistic Account. *ICSA International Conference on Speech Prosody*, Nara, Japan, 2004

[8] Holmberg, E., Hillman, R.E. and Perkell, J.S. Glottal air flow and transglottal pressure measurements for male and female speakers in soft, normal and loud voice, *J.Acoust.Soc.Am.* 84, pp. 511-529, 1988.

[9] Ní Chasaide, A. and Gobl, C. "Voice source variation". In (eds.) Hardcastle, W. and Laver, J. *The Handbook of Phonetic Sciences*. Blackwell Publishers. 1997.

[10] Hanson, H.M. and Chuang, E.S. "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data". *J.Acoust.Soc.Am.* 106(2), pp. 1067-1077, 1999.

[11] Godfrey, J., E. Holliman, and J. McDaniel. "Telephone Speech Corpus for Research and Development". *Proc. the International Conference on Acoustics, Speech, and Signal Processing*. San Francisco, CA. 1992

[12] Taylor, P. Analysis and synthesis of intonation using the tilt model. J.Acoust.Soc.Am. 107(3): 1697-1714, 2000.

[13] Boersma, P. "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampling sound". In *Proc. of Institute of Phonetic Sciences, University of Amsterdam* 17, 1993.

[14] Boersam, P. and Weenink, D. "Praat: doing phonetics by computer" (Version 4.3.04) [Computer Program]. Retrieved March 8, 2005. <http://www.praat.org>

[15] Chang, C.-C. and C.-J. Lin. "LIBSVM: a library for support vector machines". Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>