On The Impact of Non-modal Phonation On Phonological Features

M. Cernak, E. Nöth, F. Rudzicz, H. Christensen, J.R. Orozco-Arroyave, R. Arora, T. Bocklet, H. Chinaei, J. Hannink, P.S. Nidadavolu, J.C. Vásquez, M. Yancheva, A. Vann, N. Vogler

IDIAP, FAU Erlangen-Nürnberg, Uni. of Toronto, Uni. of Sheffield, Uni. de Antioquia Medellín, JHU, Intel, Stanford Uni., Uni. of California-Irvine

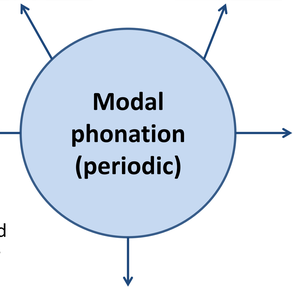
Abstract

- Voice quality in a broad sense is the characteristic auditory colouring of an individual speaker's voice.
- Such a voice quality impacts the production of the speech sounds, and we hypothesised that these changes might be captured by changes of phonological features.
- Does non-modal phonation impact phonological features?
- Yes, breathy and tense phonation impact phonological features less, creaky phonation impacts the features moderately, and harsh and falsetto phonation impact the phonological features the most.

Non-modal phonation

Creaky phonation – the arytenoid cartilages are tightly together, so that the vocal folds can vibrate only at the anterior end

Breathy phonation – there is considerable airflow in the vocal folds pulled apart, or the folds are apart only between the arytenoid cartilages.



Tense voice is produced with higher degree of overall muscular tension involved in the whole vocal tract. The higher tension of the vocal folds does not result in irregularities that are seen in harsh voice. It is characterized by richer harmonics in higher frequencies due to a less steep spectral tilt.

Falsetto voice is the most different with respect to modal voice. The voice is produced with thin vocal folds, that results in a higher pitch voice with a steeper spectral slope.

Harsh voice is a result of very high muscular tension at the laryngeal level. Pitch is irregular and low, and the speech spectrum contains more noise.

Definition Different modes of vibration of the vocal folds contribute significantly to the voice quality. The neutral mode phonation, often used in a modal voice, is one against which the other modes can be contrastively described, also called non-modal phonations.

Phonological features The Sound Patterns of English (SPE) represent a phoneme by a combination of phonological features. For example, a consonant [j] is articulated using the mediodorsal part of the tongue [+High], generated with simultaneous vocal fold vibration [+Voiced].

Are posterior probabilities of the phonological features estimated from speech changed with non-modal phonation?

Experimental setup

Design

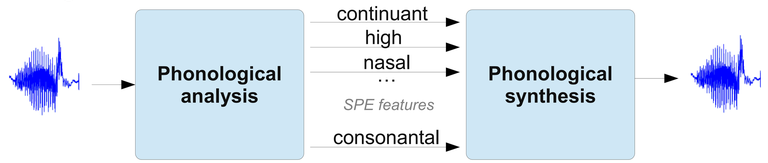


Figure: *No accepted standard system exists for describing pathological voice qualities. Qualities are labeled based on the perceptual judgments of individual clinicians, a procedure plagued by inter- and intra-rater inconsistencies and terminological confusions. Synthetic pathological voices could be useful as an element in a standard protocol for quality assessment. . .*

Analysis and synthesis

1. Feature extraction (MFCCs): converts the speech samples $\vec{x}_n$ with $n \in N$ number of frames in the speech signal into a sequence of acoustic feature observations $X = \{\vec{x}_1, \ldots, \vec{x}_n, \ldots, \vec{x}_N\}$.
2. Getting phonological posteriors: DNN (phonological analysis) converts the acoustic feature observation sequence X into a sequence of vectors $Z = \{\vec{z}_1, \ldots, \vec{z}_n, \ldots, \vec{z}_N\}$, which consists of phonological posterior probabilities $z_n^k = p(c_k|x_n)$ of K phonological features (classes) c_k.
3. Remove silence frames: $\mu_k = \frac{1}{N_s}\sum_{n=1}^{N_s} p(c_k|x_n), \forall n \Longleftrightarrow p(c_{\text{SIL}}|x_n) < 0.5$, where c_{SIL} is a posterior probability of silence class being observed, and N_S is the number of non-silence frames.
4. Calculate differential characterisation: $\Delta\mu_k = \mu_k^{\text{modal}} - \mu_k^{\text{non-modal}}$, of non-modal posteriors and modal posteriors.
5. Re-synthesize the speech signal with modal and non-modal posteriors using the phonological synthesis. The phonological synthesis was trained on Nancy (female) speech with modal phonation.

Data

- Training: The phonological analyser is trained on the Wall Street Journal (WSJ0 and WSJ1) continuous speech recognition corpora.
- Evaluation: John Laver's recordings are considered as recordings of non-modal phonation with excellent quality. The read-VQ database containing two male and two female recordings that covers five different non-modal phonations: falsetto, creaky, harshness, tense and breathiness. The recordings with modal phonation were 2.2 minutes long, and the remaining recordings with non-modal phonation were 2.0 minutes long each (i.e., altogether about 12.2 minutes of recordings).

Training procedure

- The three-state, cross-word triphone models were trained with the HTK system to get the WSJ phoneme alignments.
- 13 DNNs were trained with the Kaldi toolkit as phonological analyzers using the short segment (frame) alignment with two output labels indicating whether the phonological class exists for the aligned phoneme or not.

Results

Analysis by synthesis

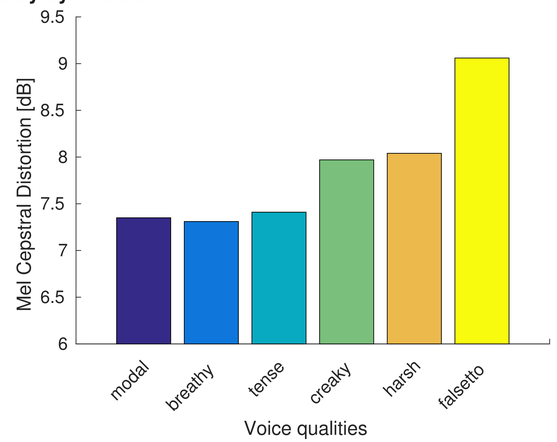


Figure: *Quality of non-modal speech synthesis measured objectivelly using Mel Cepstral Distortion in dB. The higher values indicate worse speech quality. Thus, breathy and tense phonation impact the SPE features less, creaky phonation impacts the features moderately, and harsh and falsetto phonation impact the phonological features the most.*

Overall impact of non-moda phonation

Table: *The impact of non-modal phonation on phonological features, measured as a positive (+) or negative (−) difference between the mean phonological posteriors of speech with modal phonation, and the mean phonological posteriors with non-modal phonation.*

Phonation	Invariant features	Most different features
Breathy	strident, back, voice, high	+vocalic, +tense, −nasal
Tense	strident, back, round, coronal	−low, −vocalic
Creaky	vocalic, round, high, continuant	+coronal, +conson., +nasal, −back
Harsh	strident, tense	−low, +high, −vocalic
Falsetto	strident, vocalic	+conson., +coronal, +voice, +anterior

Conclusions

We can conclude that:

1. The strident and less round and back features are more invariant features "resistant" to non-modal phonation.
2. The most impacted features for breathy and tense phonations seem to be related to vowels and nasals, creaky phonation seems to be related to both vowels and consonants, and harsh and falsetto phonations impact mostly consonants.

Idiap research institute, Martigny, Switzerland

Mail: {Milos.Cernak}@idiap.ch