Proceedings Paper:

Cernak, M., Nöth, E., Rudzicz, F. et al. (11 more authors) (2017) On the impact of non-modal phonation on phonological features. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 05-09 Mar 2017, New Orleans, LA, USA. IEEE , pp. 5090-5094. ISBN 9781509041176

https://doi.org/10.1109/ICASSP.2017.7953126

ON THE IMPACT OF NON-MODAL PHONATION ON PHONOLOGICAL FEATURES

Milos Cernak[1], *Elmar Nöth*[2*],
Frank Rudzicz[3], *Heidi Christensen*[4], *Juan Rafael Orozco-Arroyave*[5], *Raman Arora*[6],
Tobias Bocklet[7], *Brian Roark*[8], *Hamidreza Chinaei*[3], *Julius Hannink*[2], *Phani Sankar Nidadavolu*[6],
Juan Camilo Vasquez[5], *Maria Yancheva*[3], *Alyssa Vann*[9], *Nikolai Vogler*[10]

[1] Idiap Research Institute, Switzerland, milos.cernak@idiap.ch;
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, elmar.noeth@fau.de;
[3]University of Toronto, Canada; [4]University of Sheffield, UK;
[5]Universidad de Antioquia Medellín, Columbia; [6]John Hopkins University, USA; [7]Intel, Germany;
[8]Google, USA; [9]Stanford University, USA; [10]University of California-Irvine, USA

ABSTRACT

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

This paper investigates the impact of non-modal phonation on phonological posteriors, the probabilities of phonological features inferred from the speech signal using deep learning approach. Five different non-modal phonations are considered: falsetto, creaky, harshness, tense and breathiness are considered, and their impact on phonological features, the Sound Patterns of English (SPE), is investigated, in both speech analysis and synthesis tasks. We found that 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.

Index Terms— Phonological features, non-modal phonation, phonological vocoding

1. INTRODUCTION

Pathological speech is characterised by soft volume, monotone, hoarseness, breathiness, imprecise articulation and vocal tremor [1]. The project[1] titled "Analysis by Synthesis of Severely Pathological Voices", conducted at Head and Neck Surgery, UCLA School of Medicine, concluded, that *"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…"*

Even if we do not consider analysis and synthesis of pathological voices, non-modal (or aperiodic) phonation of "healthy" speakers poses challenges in current speech technology as well. For example, an American English speaker (labelled BDL) in the ARCTIC speech database [2], often used in current text-to-speech (TTS) research, happens to regularly produce creak in parts of his read sentences. This led some recent works to focus on improvements of analysis and synthesis of creaky voices [3, 4].

Recent work on non-modal phonation focuses on detection [5], analysis [6, 7] and synthesis [8] of speech with non-modal phonation. Modern computational paralinguistics tries to 1) get rid of non-modal phonation, or 2) model it, for example, for classification purposes [9]. However, the production of speech sounds with non-modal phonation has been less studied. Speech sounds can be well characterised by phonological features, and thus, we aim to study in this work the impact of non-modal phonation on phonological features. The goal is to identify the invariant, and the most impacted phonological features, and use these patterns in future work on analysis and synthesis of pathological speech.

For studying the speech with non-modal phonation, we used the read-VQ database [10], the recording of which was inspired by prototype voice quality examples produced by John Laver [11]. Five different non-modal phonations are considered: falsetto, creaky, harshness, tense and breathiness. Analysis of phonological features, the Sound Patterns of English (SPE) features [12], was performed by the PhonVoc toolkit [13]. Consequently, the inferred probabilities of the SPE features, called also phonological posteriors, were used for the re-synthesis of the speech signals. Thus, we used the analysis-by-synthesis approach to study the impact of non-modal phonation on phonological features. We analysed and re-synthesized both original Laver's recordings and the read-VQ recordings, and statistically evaluated differences on modal and non-modal phonological posteriors.

2. NON-MODAL PHONATION

We follow Laver's terminology [11] for using the term of voice quality, that is defined in a broad sense as the characteristic auditory colouring of an individual speaker's voice, and not just in a narrow sense coming from laryngeal activity. Such voice quality impacts the produces speech sounds, and we hypothesised that these changes might be captured by changes of phonological posteriors.

Different modes of vibration of the vocal folds contribute significantly to voice quality. The modal (periodic) phonation, is one against which the other modes can be contrastively described, called also non-modal (aperiodic) phonations.

Breathy and creaky voices belong to the most studied non-modal phonation types. In breathy phonation, the vibration of the vocal folds is accompanied by aspiration noise, that causes higher first formant bandwidth and missing third formant [14] due to steeper spectral tilt [15]. In creaky phonation (refereed also as vocal fry,

*Thanks to 2016 JHU workshop for funding…

[1]http://www.seas.ucla.edu/spapl/projects/pathological.html

laryngealisation), secondary vibrations occur, produced with lower fundamental frequencies. Creaky voice is a "characteristic" phonation, studied also in sociolinguistics.

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

Falsetto voice is the most opposite to modal voice [11]. The voice is produced with thin vocal folds, that results into a higher pitch voice with a steeper spectral slope.

3. EXPERIMENTAL SETUP

We use our open-source phonological vocoding platform [16] to perform phonological analysis and synthesis. Briefly, the platform is based on cascaded speech analysis and synthesis that works internally with the phonological speech representation. In the phonological analysis part, phonological posteriors are detected directly from the speech signal by Deep Neural Networks (DNNs). Binary [17] or multi-valued classification [18, 19] might be used. In the latter case, the phonological classes are grouped together based on place or manner of articulation. We followed the binary classification approach in our work, and thus each DNN determines the probability of a particular phonological class.

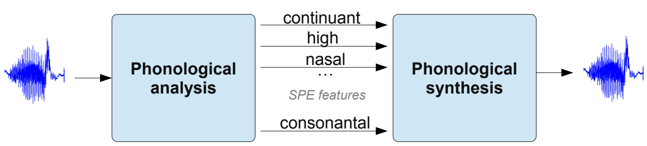


Fig. 1. Phonological analysis and synthesis.

Fig. 1 shows the phonological analysis and synthesis. We used the Sound Patterns of English (SPE) feature set [12] for training of the DNNs for phonological posterior estimation. The mapping used to map from phonemes to SPE phonological classes is taken from [20]. The distribution of the phonological labels is non-uniform, driven by mapping different numbers of phonemes to the phonological classes.

3.1. Training

To train the DNNs for phonological analysis, we first trained a phoneme-based automatic speech recognition system using mel frequency cepstral coefficients (MFCC) as acoustic features. The phoneme set comprising of 40 phonemes (including "sil", representing silence) was defined by the CMU pronunciation dictionary. The three-state, cross-word triphone models were trained with the HMM-based speech synthesis system (HTS) variant [21] of the Hidden Markov Model Toolkit (HTK) on the 90% subset of the WSJ *si_tr_s_284* set [22]. The remaining 10% subset was used for cross-validation. The acoustic models were used to get boundaries of the phoneme labels.

Then, the labels of phonemes were mapped to the SPE phonological classes. In total, 13 DNNs were trained as the phonological analyzers using the short segment (frame) alignment with two output labels indicating whether the k-th phonological class exists for the aligned phoneme or not. In other words, the two DNN outputs correspond to the target class vs. the rest. Some classes might seem to have unbalanced training data, for example, the two labels for the nasal class are associated with the speech samples from just 3 phonemes /m/, /n/, and /ŋ/, and with the remaining 36 phonemes. However, this split is necessary to train a discriminative classifier well, as all the remaining phonemes convey information about all different phonological classes. Each DNN was trained on the whole training set. The DNNs have an architecture of $351 \times 1024 \times 1024 \times 1024 \times 2$ neurons, determined empirically based on the authors' experience. The input vectors are 39 order MFCC features with the temporal context of 9 successive frames. The parameters were initialized using deep belief network pre-training done by single-step contrastive divergence (CD-1) procedure of [23]. The DNNs with the softmax output function were then trained using a mini-batch based stochastic gradient descent algorithm with the cross-entropy cost function of the KALDI toolkit [24].

Training of the phonological synthesis starts with preparing input features from the TTS database by performing the phonological analysis using the analysis DNNs. We used the Nancy database provided in the Blizzard Challenge 2011, that consits of 16.6 hours of high quality recordings of natural expressive human speech made in an anechoic chamber. The output features – modelled speech parameters – are extracted by the LPC analysis. Cepstral mean normalisation of the output features is applied before DNN training. The DNN is also initialised by pre-training, and is trained with a linear output and the mean square error cost function. The synthesis DNN is trained again with the Kaldi toolkit.

3.2. Evaluation data

We used the read-VQ database [10] in this work. Participants, 2 males and 2 females, were asked to read 17 sentences in six different phonation types: modal, breathy, tense, harsh, creaky and falsetto. The sentences were chosen from the phonetically compact sentences in the TIMIT corpus [25], four of which contained all-voiced sounds. 451 sentences were chosen in order to obtain a wide phonetic coverage, and as it is likely that it can be very difficult for speakers to maintain a constant type of phonation over a long utterance.

Participants were given prototype voice quality examples, produced by John Laver [11] and the John Kane [10], and were asked to practise producing them before coming to the recording session. For the recordings participants were asked to produce the strong versions of each phonation type and to maintain it throughout the utterance. During the recording session participants were asked to repeat the sentence when it was deemed necessary. The recordings with modal phonation were 2.2 minutes long, and the remaining recordings with non-modal phonation were 2 minutes long each (i.e., altogether about 12.2 minutes of recordings). Both Laver's and the read-VQ data were used in the evaluation.

3.3. Analysis and synthesis

Phonological analysis starts by converting speech samples $\vec{x}_n$ with $n \in N$ number of frames in the speech signal into a sequence of acoustic feature observations $X = \{\vec{x}_1, \ldots, \vec{x}_n, \ldots, \vec{x}_N\}$. Conventional cepstral coefficients can be used in this speech analysis step. Then, the analysis realised by DNNs converts the acoustic feature observation sequence X into a sequence of vectors $Z = \{\vec{z}_1, \ldots, \vec{z}_n, \ldots, \vec{z}_N\}$. The vector of phonological parameters $\vec{z}_n = [z_n^1, \ldots, z_n^k, \ldots, z_n^K]^\top$ consists of phonological posterior probabilities $z_n^k = p(c_k|x_n)$ of K phonological features (classes) c_k.

The matrix of posteriors Z thus consist of N rows, indexed by the processed speech frames, and K columns. The following analysis was done on non-silence speech frames of the evaluation data:

$$\mu_k = \frac{1}{N_s}\sum_{n=1}^{N_s} p(c_k|x_n), \forall n \Longleftrightarrow p(c_{\text{SIL}}|x_n) < 0.5, \qquad (1)$$

where c_{SIL} is a posterior probability of silence class being observed, and N_S is the number of non-silence frames. First, modal voice was analysed, followed by other non-modal phonations analysed deferentially (contrastively) to the modal voice:

$$\Delta\mu_k = \mu_k^{\text{modal}} - \mu_k^{\text{non-modal}}. \qquad (2)$$

After obtaining the SPE phonological posteriors, we used the posteriors also to re-synthesize the speech signal using the phonological synthesis. The phonological synthesis was trained on Nancy (female) speech with modal phonation, thus impacted (distorted) phonological posteriors cased by non-modal phonation, should result in lower quality re-synthesized speech.

4. RESULTS

4.1. Analysis

We evaluated first original Laver's recordings. They are considered as recordings of non-modal phonation with excellent quality, however only one utterance per the phonation type is available, and thus they are speaker-specific. Fig. 2a shows the analysis of Laver's recordings, followed by the analysis of the read-VQ evaluation data in Fig. 2b.

Table 1 lists the invariant and the most different features between speech with modal and non-modal phonations.

Table 1. *The impact of non-modal phonation on the SPE posterior features, measured as a 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, consonantal, nasal, back
Harsh	Strident, tense	Low, high, vocalic
Falsetto	Strident, vocalic	Consonantal, coronal, voice, anterior

4.2. Synthesis

We evaluated synthesized speech of 2 female speakers from the read-VQ database using the Mel Cepstral Distortion (MCD) [26] between original and synthesized speech samples. Lower MCD values indicate higher speech quality of the synthesized speech samples. Fig. 3 shows synthesis results of the read-VQ data.

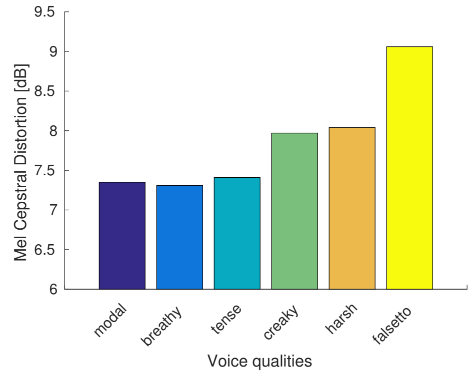


Fig. 3. Quality of non-modal speech synthesis, measured objectivelly using Mel Cepstral Distortion in dB. The higher values indicate worse speech quality.

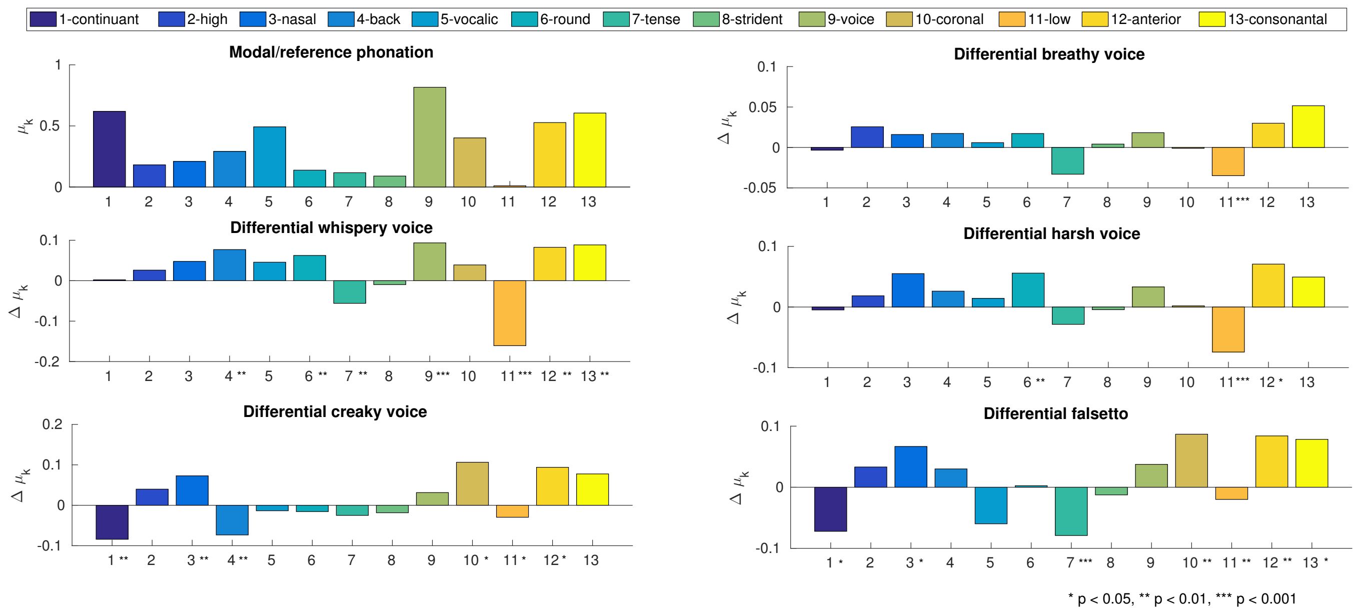
4.3. Discussion

TODO: Add discussion about these results:

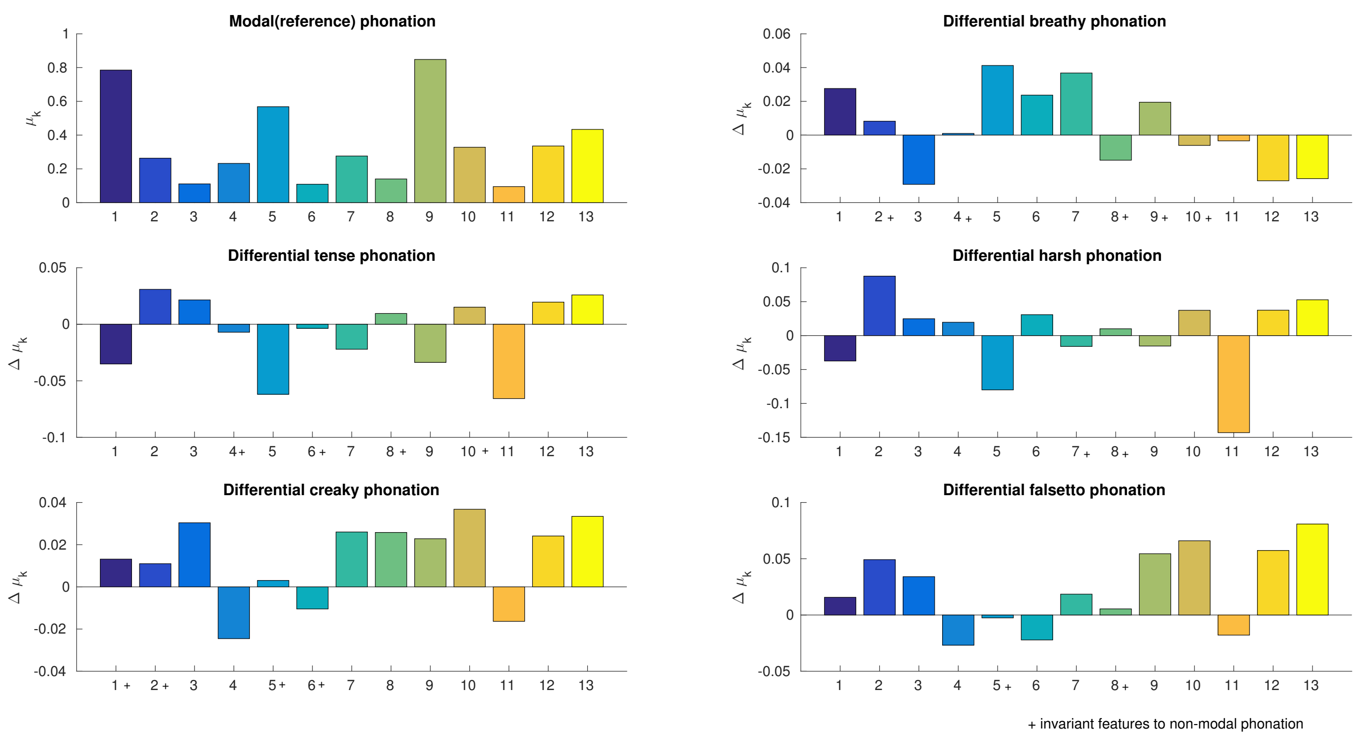
1. By visual comparison of Fig. 2a and Fig. 2b, we can conclude that the impact of non-modal phonation on phonological posteriors is roughly similar for both the Laver's and read-VQ recordings. Invariant phonological features were estimated from the read-VQ analysis that is speaker-independent.
2. As shows Fig. 3, 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.
3. As lists Tab. 1, strident, and less round and back features, are more invariant features "resistant" to non-modal phonation, the rest of the features is heavily impacted. The most impacted features for breathy and tense phonations seem to be related to vowels (such as low and nasal), creaky phonation seems to be related to both vowels and consonants (such as coronal and nasal), and harsh and falsetto phonations impact mostly consonants (coronal, anterior, consonantal).
4. Linking of these findings to pathological speech...

5. ACKNOWLEDGEMENTS

This work has been conducted with the support of the 2016 JHU workshop.



(a) Analysis of the Laver's recordings. The stars next to the indices of the phonological classes indicate statistical significance of difference between the modal and particular non-modal phonation.



(b) Analysis of the read-VQ recordings. The plus next to the indices represent the invariance (where statistical significance of differences is $p > 0.001$), and the rest of the indices represent statistically significant ($p < 0.001$) differences between the modal and particular non-modal phonation.

Fig. 2. Mean modal SPE posteriors μ_k (top-left figures) and differentials $\Delta\mu_k$ of non-modal phonations with respect to the modal voice.

6. REFERENCES

[1] Rajesh Pahwa and Kelly E. Lyons, Eds., *Handbook of Parkinson's Disease*, CRC Press Book, fourth edition edition, Mar. 2007.

[2] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223 – 224.

[3] Tamás G. Csapó and Géza Németh, "Modeling Irregular Voice in Statistical Parametric Speech Synthesis With Residual Codebook Based Excitation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 209–220, Apr. 2014.

[4] Tamás G. Csapó, Géza Németh, Milos Cernak, and Philip N Garner, "Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in *Proc. of EUSIPCO*, 2016.

[5] Thomas Drugman, John Kane, and Christer Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech & Language*, vol. 28, no. 5, pp. 1233–1253, Sept. 2014.

[6] Nicolas Malyska, *Analysis of nonmodal glottal event patterns with application to automatic speaker recognition*, Ph.D. thesis, Harvard University – MIT Division of Health Sciences and Technology, USA, 2008.

[7] Nicolas Malyska, Thomas F. Quatieri, and Robert B. Dunn, "Sinewave Representations of Nonmodality," in *Proc. of Interspeech*, 2011, pp. 69–72.

[8] Philbert Bangayan, Christopher Long, Abeer A. Alwan, Jody Kreiman, and Bruce R. Gerratt, "Analysis by synthesis of pathological voices using the Klatt synthesizer," *Speech Communication*, vol. 22, no. 4, pp. 343–368, Sept. 1997.

[9] Björn Schuller and Anton Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, Wiley, Nov. 2013.

[10] John Kane, *Tools for analysing the voice*, Ph.D. thesis, Trinity College Dublin, Dublin, Sept. 2012.

[11] John Laver, *The Phonetic Description of Voice Quality*, Cambridge Studies in Linguistics. Cambridge University Press, Mar. 2009.

[12] N. Chomsky and M. Halle, *The Sound Pattern of English*, Harper & Row, New York, NY, 1968.

[13] Milos Cernak and Philip N Garner, "PhonVoc: A Phonetic and Phonological Vocoding Toolkit," in *Proc. of Interspeech*, San Francisco, CA, USA, 2016.

[14] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers.," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, Feb. 1990.

[15] H. M. Hanson, "Glottal characteristics of female speakers: acoustic correlates.," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 466–481, Jan. 1997.

[16] Milos Cernak, Afsaneh Asaei, Pierre-Edouard Honnet, Philip N. Garner, and Hervé Bourlard, "Sound Pattern Matching for Automatic Prosodic Event Detection," in *Proc. of Interspeech*, 2016.

[17] Dong Yu, Sabato Siniscalchi, Li Deng, and Chin-Hui Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *Proc. of ICASSP*. March 2012, IEEE SPS.

[18] F. Stouten and J.-P. Martens, "On The Use of Phonological Features for Pronunciation Scoring," in *Proc. of ICASSP*. May 2006, vol. 1, p. I, IEEE.

[19] Ramya Rasipuram and Mathew Magimai.-Doss, "Integrating articulatory features using Kullback-Leibler divergence based acoustic model for phoneme recognition," in *Proc. of ICASSP*. May 2011, pp. 5192–5195, IEEE.

[20] Milos Cernak, Stefan Benus, and Alexandros Lazaridis, "Speech vocoding for laboratory phonology," 2016.

[21] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda, "The HMM-based Speech Synthesis System Version 2.0," in *Proc. of ISCA SSW6*, 2007, pp. 131–136.

[22] Douglas B. Paul and Janet M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, Stroudsburg, PA, USA, 1992, HLT '91, pp. 357–362, Association for Computational Linguistics.

[23] Geoffrey E. Hinton, Simon Osindero, and Yee W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[24] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *Proc. of ASRU*. Dec. 2011, IEEE SPS, IEEE Catalog No.: CFP11SRW-USB.

[25] Linguistic Data Consortium, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993.

[26] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. of ICASSP*. May 1993, vol. 1, pp. 125–128 vol.1, IEEE.