Exploiting Vocal-Source Features to Improve ASR Accuracy for Low-Resource Languages

Raul Fernandez[1], *Jia Cui*[1], *Andrew Rosenberg*[2], *Bhuvana Ramabhadran*[1], *Xiaodong Cui*[1]

[1]IBM TJ Watson Research Center, Yorktown Heights, NY – USA
[2]Queens College (CUNY), Flushing, NY – USA

{fernanra,jiacui,bhuvana,cuix}@us.ibm.com, andrew@cs.qc.cuny.edu

10.21437/Interspeech.2014-206

Abstract

A traditional framework in speech production describes the output speech as an interaction between a source excitation and a vocal-tract configured by the speaker to impart segmental characteristics. In general, this simplification has led to approaches where systems that focus on phonetic segment tasks (e.g. speech recognition) make use of a front-end that extracts features that aim to distinguish between different vocal-tract configurations. The excitation signal, on the other hand, has received more attention for speaker-characterization tasks. In this work we look at augmenting the front-end in a recognition system with vocal-source features, motivated by our work with languages that are low in resources and whose phonology and phonetics suggest the need for complementary approaches to classical ASR features. We demonstrate that the additional use of such features provides improvements over a state-of-the-art system for low-resource languages from the BABEL Program.
Index Terms: ASR features, vocal source features

1. Introduction

A common model in speech processing describes the acoustic waveform as the result of the interaction between an excitation signal generated in the glottis with a vocal tract configured by the speaker to confer the acoustic characteristics of the phonetic segments of the language. Under this paradigm, speech processing tasks that focus on the segmental component of the language (e.g., phonetic transcription) make use of front-end processors that focus on extracting features related to the vocal tract (e.g., Mel Cepstral or Perceptual Linear Prediction coefficients). The excitation signal, on the other hand, is generally considered to be minimally influenced by the text (with the exception of the fundamental frequency of vibration $F0$, and its role in signaling lexical tone in tone languages), but is known to play a role in conveying the timbral characteristics of speech and to be employed communicatively to transmit paralinguistic information. Properties of the excitation that quantify more than just the fundamental have consequently been exploited for tasks such as speaker identification [1], affect recognition [2], as well as in the description and diagnosis of pathological voice conditions [3], but have received less attention for speech recognition applications.

It is known, however, that the excitation signal is used to generate phonological contrasts in many languages (such as the distinction between breathy- and modal-phonation consonants in languages like Hindi or Bengali). More generally, phonetic features of the excitation may help characterize sounds whose production involve the glottis and vocal chords in prominent roles (such as the production of aspirated or ejective consonants) or sounds whose generation co-occur with changes in voice-quality (such as the laryngealization observed in the generation of low tones in tonal languages). We are interested in the use of these types of features to complement the standard front-end pipeline of classical "vocal-tract" features within the acoustic model of a recognition system. Our interest is motivated by our recent work with languages in the BABEL program, which seeks to develop methods to build speech recognition technology for a much larger set of languages than previously investigated. This interest is twofold: many of these languages have not been studied within the ASR community for too long, and their phonetic and phonological properties are distinct enough from better understood languages, that their automatic recognition may necessitate or benefit from alternative approaches. The second is the small amount of resources available for such languages: we hope that the use of these alternative features may also provide a certain amount of phonetic redundancy to help with phonetic disambiguation in light of the data sparsity issue. In this paper we focus on two languages from the BABEL program, Zulu and Lao, that meet these criteria. Zulu is a Bantu language whose phonology includes the use click consonants; contrastive aspirated, implosive and depressor consonants; and word-level tone. Lao is a heavily tonal language, which deploys 6 contrastive lexical tones, and also makes use of contrastive aspiration. The paper is structured as follows: In Section 2 we review a series of features from the voice-source analysis and voice-quality literature that we are applying to the recognition task. Section 3 describes the data sources and systems evaluated, and in Section 4 we offer some concluding remarks.

2. Vocal Source Features

In this section, we review various features proposed in the glottal source analysis literature. For the experiments reported in Section 3 we have relied extensively on the software implementation of these features available and distributed as part of the COVAREP repository[4].

2.1. Sum of Residual Harmonics

The Sum of Residual Harmonics (SRH) was introduced as a robust criterion to estimate the fundamental frequency of a speech signal [5]. If $e(t)$ denotes the inverse-filtered LPC residual (of order 10 in the implementation), and $E(f)$ the spectrum of a Hanning-windowed segment long enough to accommodate a few speech cycles, then the frequency-dependent SRH criterion is defined as:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{max}} [E(kf) - E((k - \frac{1}{2})f)], \quad (1)$$

where N_{max} denotes the maximum number of harmonics taken into account. Eqn. 1 can be searched in a frequency range $[f0_{min}, f0_{max}]$ to yield an estimate of the fundamental frequency $f0 = \text{argmax}_f\, SRH(f)$, for which the SRH parameter is then defined as

$$SRH^* = SRH(f0) \tag{2}$$

2.2. Maxima Dispersion Quotient

The Maxima Dispersion Quotient (MDQ) is a pitch-synchronous parameter that characterizes how impulse-like the glottal excitation is, and which has been shown to help distinguish between various voice quality types [6]. Based on the zero-phase, dyadic wavelet analysis of the LPC-inverse-filtered residual signal $e(t)$ across 7 bands,

$$y_i(t) = e(t) * \phi\left(\frac{t}{2^i}\right) \quad i = 0, \cdots, 6 \tag{3}$$

where $\phi(t)$ is the cosine-modulated Gaussian mother-wavelet:

$$\phi(t) = -\cos(2\pi f_N t)\exp\left(-\frac{t^2}{2\tau^2}\right), \tag{4}$$

f_N is the Nyquist frequency, and $\tau = \frac{1}{2}f_N$, MDQ is then defined for the n^{th} glottal cycle as

$$MDQ_n = \frac{\frac{1}{K}\sum_{i=0}^{K-1}|GCI^n - m_i|}{T_0^n} \tag{5}$$

where T_0^n is the fundamental period of the n^{th} glottal cycle, GCI^n is that cycle's glottal closure instant, and m_i is the location of the maximum of y_i (the maxima over the bands are searched over an interval of length $0.4T_0^n$ centered at GCI^n). Eqn. 5 measures how much, on the average, the maxima of the wavelet-filtered signals deviate from the glottal closure instant. For an ideal Dirac delta pulse, MDQ is zero, and, as shown in [6], the location of m_i are more closely clustered around GCI^n for the "sharper" excitations of tense voice and more scattered for the excitations associated with breathy voice (leading to lower and higher MDQ, respectively).

2.3. Peak Slope

The Peak Slope (PS) parameter [7] is related to the method presented in 2.2. It too is based on a multi-band dyadic decomposition using the wavelet $\phi(t)$ in Eqn. [4]. In this implementation, however, the input to the convolution in Eqn. [3] is the actual speech signal, and the analysis is carried out at a constant frame rate (using a 40 msec. window length and a 10 msec. shift). For each frame, and each band i, the peak amplitude A^i of the absolute value of the filtered signal is found and paired with the center frequency of the i^{th} filter. PS then provides a measure of spectral tilt by looking at the decay of the A^i values, and is defined as the slope of the linear fit to the set of points $\left\{\frac{f_N}{2^i}; \log_{10} A^i\right\}_{i=0}^{6}$, where f_N is the Nyquist frequency.

2.4. Normalized Amplitude Quotient

The Normalized Amplitude Quotient (NAQ) is a pitch-synchronous time-domain parameter introduced in [8] as a measure of the closed-phase quotient (the portion of a cycle during which the glottis is closed) and which is related to the decay (smooth or abrupt) of the glottal pulse during the open-to-closed phase transition. It has been shown to be more robust under noisy conditions than direct estimates of the closed-phase quotient that require explicit identification of the open and closed phases of a glottal cycle, by relying only on extrema extractable from the glottal flow waveform and its derivative. If $g(n)$ is a cycle of period T_0 from the glottal flow waveform reaching a maximum value f_{max}, and $g'(n)$ is its derivative reaching a minimum value $-d_{peak}$, then NAQ is defined as

$$NAQ = \frac{1}{T_0}\frac{f_{max}}{d_{peak}} \tag{6}$$

2.5. Quasi Open Quotient

The Open Quotient (OQ) is defined as the portion of a glottal cycle during which the glottis remains open. The Quasi Open Quotient (QOQ) is a variant of this definition that follows a similar approach to the estimation of the previous parameter in that it introduces an approximation with a feature that is a function of the glottal flow extrema, and which has been shown to reliably distinguish between different phonation types [9]. It is a pitch-synchronous parameter, and defined as the portion of a glottal cycle (of period T_0) during which the glottal flow remains above a threshold (here 50% of the maximum flow). Centered around the time instant T_{max} at which $g(t)$ reaches its maximum f_{max}, an interval $(T_1, T2)$ is grown, with $T_1 < T_{max} < T_2$, to satisfy the above condition. Then QOQ is defined as

$$QOQ = \frac{T_2 - T_1}{T_0} \tag{7}$$

2.6. First-to-Second-Harmonic Difference

The first-to-second-harmonic difference (H_{12}) is a pitch-synchronous spectral parameter that has been widely used in the literature to characterize vocal source differences. It has been shown, e.g., to correlate with variations in OQ [10]. It is computed from the glottal flow derivative waveform. Given a glottal cycle of period T_0, a GCI-centered Hanning window of length $3T_0$ is used to compute the short-time spectrum, which is used to identify the locations and amplitudes (in dB) of the first two harmonic peaks H_1 and H_2. The parameter is defined as

$$H_{12} = H_1 - H_2 \tag{8}$$

2.7. Harmonic Richness Factor

The Harmonic Richness Factor (HRF) [11] is a pitch-synchronous parameter that tries to quantify the amount of harmonic information in the spectrum of the glottal source, and is defined as the ratio between the sum of the harmonic amplitudes and the amplitude of the fundamental. It can be estimated following a similar procedure to the estimation of H_{12} by extracting K harmonics and their amplitudes H_i from the short-time spectrum of the glottal flow derivative (where K is the number of harmonics found in the $(f_0 - 5kHz)$ range in the current implementation) , and then computing:

$$HRF = \frac{\sum_{k=1}^{K} A_i}{A_0} \tag{9}$$

2.8. Parabolic Spectral Parameter

The Parabolic Spectral Parameter (PSP) [12] is a frequency-domain parameter extracted from the spectrum of the glottal flow derivative waveform $g'(n)$computed pitch-synchronously

over one glottal cycle (which has been scaled to have unit energy), and which attempts to quantify the spectral decay over the low frequencies. If $X(f)$ denotes the spectrum of $g'(n)$ thus computed, then the PSP parameter is derived under the assumption that a parabolic function $\hat{X}(f) = a_2 f^2 + a_0$ provides a good fit to the spectrum. In this case, a_2 is used to describe the shape of the decay and can be estimated by minimizing the following normalized error criterion:

$$NE = \frac{\sum_{f=0}^{N-1}(X(f) - a_2 f^2 + a_0)^2}{\sum_{f=0}^{N-1} X(f)^2} \tag{10}$$

The range N is iteratively grown (starting from a value of 3), and a_2 refit, until NE exceeds a threshold (0.01 in this implementation). PSP is then defined as

$$PSP = \frac{a_2}{a_{2_{max}}} \tag{11}$$

where a_2 is the fit from the procedure just described for a glottal cycle of period T_0, and $a_{2_{max}}$ is the corresponding fit to a DC signal of unity amplitude and length T_0 (providing the maximal decay of a hypothetical DC flow). See [12] for more details.

2.9. Probability of Creak

The probability of creak used in this paper is the implementation available as part of the GLOAT toolbox [4], and is based on a machine-learning approach that uses a single-layer Neural Network (NN) classifier to predict a probability of creak from a 36-dimensional input feature vector. This approach is corpus independent in the sense that the NN model has been pre-trained on various speech materials containing hand-labeled creak annotations (representing various speakers, genders, languages, and speaking styles), and no further training or adaptation to the target corpus takes place. The input features to the NN consist of a base 12-dimensional vector (plus its first and second derivatives) containing measures related to f0, energy, spectral power, first-to-second-harmonic difference, zero-crossing rate, glottal pulse-to-pulse similarity, and intra-frame periodicity reported in [13] and [14] for creak detection.

2.10. Glottal Shape Parameter R_d

In order to characterize the glottal flow waveform, various parametric models have been proposed, one of the most commonly used among which is the Liljencrants-Fant (LF) model of the glottal flow derivative. For a glottal cycle of known period T_0, the LF glottal flow derivative is given by

$$g'(n) = \begin{cases} \dfrac{-E_e e^{\alpha(t-T_e)} \sin\left(\frac{\pi t}{T_p}\right)}{\sin\left(\frac{\pi T_e}{T_p}\right)} & 0 \le t \le T_e \\ \frac{E_e}{\epsilon T_a} e^{\epsilon(T_e - t)} - e^{\epsilon(T_e - T_c)} & T_e < t < T_c \\ 0 & T_c < t < T_0 \end{cases} \tag{12}$$

where T_e and T_c denote the beginning and end of the closed phase portions of the cycle, respectively. In practice we can assume that $T_c = T_0$ [15]. The ϵ parameter may not be considered an independent parameter since it can be shown that for small enough T_a, $\epsilon T_a \approx 1$. Furthermore, the value of α is derivable from the remaining parameters from the constraint $g'(T_p) = 0$ since T_p is the instant at which the glottal flow reaches its maximum (i.e. the derivative is zero). With these constraints, the shape of a glottal pulse is fully determined by a four-parameter vector $\theta = [T_a, T_e, T_p, E_e]$, of which the first three describe the shape of the pulse, and E_e the amplitude. A more established and equivalent formulation of the model describes the shape of the pulse in terms of the following 3 parameters: $R_a = \frac{T_a}{T_0}$, $R_g = \frac{T_0}{2T_p}$, and $R_k = \frac{T_0 - T_p}{T_p}$. An analysis of the way in which these 3 parameters covary for various voice quality types led to the formulation of a single shape parameter, known as R_d, which can be approximated from the others as [16]

$$R_d \approx (1/0.11)(0.5 + 1.2R_k)(R_k/4R_g + R_a). \tag{13}$$

Conversely, simple regression formulae have been provided to predict the R_* parameters from R_d [16].

A pitch-synchronous estimate of the R_d parameter is used as a vocal source feature in this study. The estimation procedure involves deriving a residual all-pass transfer function, and introducing a phase-distortion criterion as a function of R_d that can be efficiently minimized (the details of the derivations are omitted due to length constraints, but are described in [17, 18], and implemented in [4])

2.11. Fundamental Frequency Variation

Fundamental Frequency Variation (FFV) is a robust frame-based parameter proposed to measure instantaneous changes in fundamental frequency which is computed without having to explicitly calculate an F0 contour or inter-frame differences, and which, unlike F0, is defined for all frames regardless of their voicing status [19]. It is estimated by calculating the normalized dot product between 2 FFT spectra across a short region at different dilations of one or both spectra. The intuition is that the spectrum will dilate over time with increasing $F0$ (i.e., the space between harmonics will increase). On the other hand, if pitch is decreasing the harmonics will contract. The FFV spectrum is a measure of this dot product (the magnitude of similarity) at a range of values of a dilation parameter, τ.

The FFV calculation involves an external window, an internal window, and a window separation parameter w_{sep}. The two FFT spectra are calculated from two asymmetric windows before and after a time, t_i. The window functions are defined by an external (away from t_i) half Hamming window and internal (toward t_i) half Hanning window with its peak at $t_i \pm w_{sep}/2$. Identifying the instant τ^* at which the FFV spectrum reaches its maximum provides a measure of instantaneous F0 variation. However, it is more common to apply a filter bank of five or seven filters, whose frequency response shapes are designed to pick up different patterns of variation, to yield a frame-based vector measure of changes in $F0$. These filters are designed to have responses that correspond to stable pitch (the center filter), with increasing degrees of change in pitch as you move away from the center filter (see [19] for more details).

Recently a number of groups have demonstrated the utility of FFV as an input feature for speech recognition tasks [20], which motivates our inclusion with the source features previously described. For the experiments reported here, we have made used the software implementation of FFV available at [21] using a 15-msec. external window, a 11-msec. internal window, $w_{sep} = 10$ msecs., and a 7-filter bank.

3. System Descriptions and Evaluation

The Limited Language Pack for the Babel option period includes the languages Zulu and Lao. The training data for each of these languages consists of 20 hours of audio containing about 10 hours of conversational speech. To exploit the use of these

features within a ASR system, we built a baseline Deep Neural Network (DNN) system. The setup for hybrid decoding that uses the posteriors from the DNN is exactly the same as the one described in earlier work [22]. The construction of the baseline feature set is as follows: 13-dimensional mean-normalized perceptual linear prediction (PLP) features with vocal tract length normalization (VTLN) are used as the primary acoustic features, extracted with a 10 msec. frame skip. An acoustic context (CTX) of 4 frames on either side, and linear discriminant analysis (LDA) is then used to project the resulting dimensionality of the spliced 9-frame vector down to 40. The LDA features are further decorrelated by a global semi-tied covariance (STC) [23] matrix on top of which speaker-adapted training (SAT) using feature-space maximum-likelihood linear regression (FMLLR) (i.e. constrained MLLR (cMLLR)) [24] is applied to reduce the speaker variability. Finally, 9 frames of FMLLR features are concatenated together to form a 360-dimension vector that serves as the input features to the DNNs.

We define the set of Vocal Source Features (VSF) to be the 10-dimensional vector comprising the features described in sections 2.1- 2.10. All the pitch-synchronous parameters are converted to a frame-level representation, in order to be able to merge them with the standard front-end frames, as follows: parameters are paired with their corresponding glottal-closure instants, and interpolated over the length of the input to obtain a signal containing a "continuous" evolution of the parameters as a function of time. This interpolated waveform is then converted to frames by averaging 25 msec. windows with a skip of 10 msec. All glottal-closure instants are determined using the algorithm in [25] (with the implementation distributed in [4]).

To be able to assess the contributions of different types of features, we start by doing a late merge of the VSF features into our best feature set to date, which consisted of a system built with PLP and FFV features merged prior to the FMLLR operation. Specifically, the following 4 systems are evaluated:

1. the baseline system including only PLP-based inputs;
2. a system where PLP and FFV features are merged prior to FMLLR (i.e., FMLLR(PLP+FFV),
3. an extension of the system in (2) where the VSF features are merged with FMLLR(PLP+FFV) (i.e., VSF+FMLLR(PLP+FFV)),
4. a variation of (3) where F0 is also appended to the VSF set (i.e., VSF+F0+FMLLR(PLP+FFV))

where, for completeness, we also report on a system (1) built without FFV features, and one (4) where the effect of F0, which is not a novel investigation in this paper, is also assessed.

The DNN acoustic models are composed of 5 hidden layers with sigmoid activations, with each hidden layer consisting of 1024 hidden units. There are 1000 units in the softmax output layer. The DNN models are initialized by a layer-wise discriminative pre-training [26]. After this initialization, they are optimized by 14 iterations of cross-entropy (CE) training. The language model (LM) is a modified Knesser-Ney Smoothed bigram LM for Zulu, and a trigram LM for Lao built on the acoustic transcripts available for the Limited Language Pack. The lexicon sizes for Zulu and Lao are 15.9K and 3.6K respectively.

Table 1 shows the word-error rate (WER) for the 4 systems and 2 languages described, as evaluated on each language's 20-hour development set. We can see that the use of the 10-dimensional VSF features investigated (System 3) reduces the error by 0.8% absolute from the previous best performance (System 2) attained for these languages. That system, in turn, had been able to improve by 1.3% and 2.8% for Zulu and Lao respectively, by adding the 7-dimensional FFV representation to a PLP-based input. The effect of explicitly using $F0$ (System 4) is inconclusive: although we gain a further 0.2% reduction for Zulu, the WER starts to degrade for Lao after including this feature (in spite of the lexical-tone property of Lao). It is unclear at this stage why that is, but both of the experiments corroborate the hypothesis that further reductions in WER are possible by looking at source features that go beyond F0.

Table 1: *Development-set WER for the various systems and languages.*

	Zulu	Lao
(1)=Baseline	78.9%	67.1%
(2)=(1)+FFV	77.6%	64.3%
(3)=(2)+VSF	76.8%	**63.5%**
(4)=(3)+F0	**76.6%**	63.6%

4. Conclusions

In this work we have reported on initial experiments that use features derived from the vocal excitation to improve the WER on low-resource languages with distinct phonological and phonetic characteristics from traditional ASR languages. We have used features that have received attention in the voice-quality literature, but whose use within ASR tasks remains novel, and demonstrated a reduction of 0.8% for Zulu and Lao over a previous FFV-augmented system (compounded with these previously reported FFV features, the gains are 2.1% and 3.6% respectively). Although we have shown effectiveness in a limited resource context, much work remains to be done to bring these WER figures within the ranges of state-of-the-art systems for majority languages. In the future, we would like to (i) explore whether these features provide a consistent improvement over our baseline features as the amount of training data increases; (ii) expand on this initial proposal and look at other glottal source features; (iii) explore estimation methods that are robust to noisy channels, telephony bandwidth; (iv) explore other variants of the acoustic modeling tested, by merging the VSF features earlier into the modeling pipeline; and (v) exploit their use within Gaussian-Mixture Models. Finally, although we have initially evaluated the proposal on 2 languages from the BABEL program, we would like to validate it on other languages, like Bengali and Vietnamese, where we believe the approach may be well-motivated by their linguistic characteristics.

5. Acknowledgments

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. This effort uses the IARPA Babel Program language collection releases IARPA-babel203b-v3.1a and IARPA-babel206b-v0.1e limited language packs. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

6. References

[1] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech and Audio Proc.*, vol. 7, no. 5, pp. 569–586, 1999.

[2] R. Fernandez and R. W. Picard, "Recognizing affect from speech prosody using hierarchical graphical models," *Speech Communication*, vol. 53, no. 9-10, pp. 1088–1103, 2011.

[3] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar-Biarge, V. Nieto-Lluis, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, and J. I. Godino-Llorente, "Glottal source biometrical signature for voice pathology detection," *Speech Communication*, vol. 51, no. 9, pp. 759–781, 2009.

[4] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP–A collaborative voice analysis repository for speech technologies," in *ICASSP*, 2014.

[5] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Interspeech*, 2011, pp. 1973–1976.

[6] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *Computer Speech and Language*, vol. 21, no. 6, pp. 1170–1179, 2002.

[7] J. Kane and G. C., "Identifying regions of non-modal phonation using features of the wavelet transform," in *Interspeech*, 2011, pp. 177–180.

[8] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. of America*, vol. 112, no. 2, pp. 701–710, 2002.

[9] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Interspeech*, 2007, pp. 1410–1413.

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

[11] D. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[12] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter – a new method for quantification of the glottal flow," *Speech Communication*, vol. 22, no. 1, pp. 67–79, 1997.

[13] K. J. and T. Drugman, "Improved automatic detection of creak," *Computer Speech and Language*, vol. 27, no. 4, pp. 1028–1047, 2013.

[14] C. Ishi, K. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 16, no. 1, pp. 47–56, 2011.

[15] G. Fant, J. Liljencrats, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[16] G. Fant, "The LF model revisited. Transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[17] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 19, no. 5, pp. 1080–1090, 2011.

[18] ——, "Function of phase-distortion for glottal model estimation," in *ICASSP*, 2011, pp. 4608–4611.

[19] K. Laskowski, J. Edlund, and M. Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems," in *ICASSP*, 2008, pp. 5041–5044.

[20] F. Metze, Z. A. W. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen, and V. H. Nguyen, "Models of tone for tonal and non-tonal languages," in *Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 261–266.

[21] K. Laskowski, "Fundamental Frequency Variation (FFV): A normative implementation in C," http://www.cs.cmu.edu/~kornel/software_ffv.html, 2009.

[22] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained Deep Neural Networks to large vocabulary speech recognition," in *Interspeech*, 2012.

[23] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 3, pp. 272–281, 1999.

[24] ——, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[25] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Communication*, vol. 55, no. 2, pp. 295–314, 2013.

[26] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Improving training time of deep belief networks through hybrid pre-training and larger batch sizes,," in *Proc. NIPS Workshop on Log-linear Models*, 2012.