An Unsupervised Approach to Glottal Inverse Filtering

Sayan Ghosh, Eugene Laksana
Institute for Creative Technologies
University of Southern California
Los Angeles, California
Email: {sghosh,elaksana}@ict.usc.edu

Louis-Philippe Morency
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania
Email: morency@cs.cmu.edu

Stefan Scherer
Institute for Creative Technologies
University of Southern California
Los Angeles, California
Email: scherer@ict.usc.edu

***Abstract*—The extraction of the glottal volume velocity waveform from voiced speech is a well-known example of a sparse signal recovery problem. Prior approaches have mostly used well-engineered speech processing or convex L_1-optimization methods to solve the inverse filtering problem. In this paper, we describe a novel approach to modeling the human vocal tract using an unsupervised dictionary learning framework. We make the assumption of an all-pole model of the vocal tract, and derive an L_1 regularized least squares loss function for the all-pole approximation. To evaluate the quality of the extracted glottal volume velocity waveform, we conduct experiments on real-life speech datasets, which include vowels and multi-speaker phonetically balanced utterances. We find that the the unsupervised model learns meaningful dictionaries of vocal tracts, and the proposed data-driven unsupervised framework achieves a performance comparable to the IAIF (Iterative Adaptive Inverse Filtering) glottal flow extraction approach.**

I. Introduction

Glottal inverse filtering is a very important field in speech processing, with applications as diverse as speech compression, synthesis and recognition of paralinguistic attributes such as emotion and voice quality [1]. The process of inverse filtering involves an understanding of the speech production model, particularly the production of voiced sounds. In this process the glottis shapes a constant airflow input to produce a train of pulses $G(z)$ during voiced sound generation, which is called the glottal volume velocity waveform. Periodicities in the airflow stream are produced by the opening/closing of the glottis.

When the glottis closes, the glottal volume velocity airflow resonates in the vocal tracts, leading to voiced sounds being produced. The source filter model assumes that the sound source and the vocal tract are independent. The final stage in speech generation is the impedance $L(z)$ created by the lip radiation.

The vocal tract is modeled as a linear filter $V(z)$ and the glottal excitation signal $G(z)$ is estimated from the residual that is the non-linear part of the speech signal. An assumption commonly followed in the literature [2] is the all-pole model of the vocal tract, where the vocal tract $V(z)$ is modeled as :

$$V(z) = \frac{1}{1 - \sum_{p=1}^{P} a_p z^{-p}} \tag{1}$$

The glottal flow excitation can be measured using a laryngograph, however it is possible to separate the source and the vocal tract filter and estimate the excitation through a computational approach. This is a difficult deconvolution problem, for which various methods have been proposed in the literature [1] [3]. Most of these proposed approaches are well-engineered, with an additional estimation of the parameters such as the pitch, glottal opening/closure instants (such as pitch synchronous IAIF [4]). Additionally, estimation of vocal tract and glottal flow waveforms are generally done on limited data such as single frames of speech.

In this paper, we investigate glottal inverse filtering as an unsupervised learning problem, where the vocal tract parameters, along with the basis dictionary atoms and excitation signals are jointly estimated from a large corpus of speech from one or multiple speakers. Our proposed approach is motivated by the success of sparse coding and deconvolutional networks, which have been applied to fields such as feature extraction for object recognition [5], and image denoising [6].

Our primary research questions discussed in this paper are:
Q1. Is it possible to construct a data-driven unsupervised learning framework for glottal inverse filtering with minimal prior assumptions on the speech production model?
Q2. Are the glottal volume velocity waveforms extracted from the proposed model comparable to those extracted from state-of-the-art inverse filtering approaches on continuous and real-life speech data?
Q3. Do vocal tract dictionaries generalize in a speaker-independent manner across different voice quality categories such as breathy, modal and tense voices ?
We summarize the remainder of our paper as follows - in Section II, we discuss prior work, and in Section III we describe our proposed unsupervised framework for glottal inverse filtering. We conduct experiments on real-life speech datasets and evaluate the performance of our approach in Sections IV and V, concluding the paper in Section VI.

II. Related Work

The classic approach for glottal inverse filtering is LPC (Linear Predictive Coding) based estimation [7], where various phases in the glottal excitation, such as glottal closure and opening instants are estimated from an analysis

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

of the LPC residual. Iterative Adaptive Inverse Filtering (IAIF) [4] was proposed by Alku et al., in which the glottal excitation waveform is estimated in an iterative filtering process by first canceling the effects of lip radiation and estimating a lower-order vocal tract model, after which the glottal excitation is obtained by inverse filtering with a higher-order model. Recently, various approaches [8] have been proposed in which the L_1 sparsity of the residual is optimized directly, based on a linear constraint. Bayesian methods have also been introduced, such as [9] in which the block sparsity of the glottal flow is encoded using a prior, and [10], [11], where Bayesian priors and compressive sensing are respectively investigated in a TVLP (Time Varying Linear Prediction) framework. Scherer et al. [12] proposed to estimate the OQ (open quotient) in the glottal excitation using an ANN (Artificial Neural Network) and compared it to other approaches for estimating OQ. Airiksinen et al. [13] used a DNN (Deep Neural Network) to estimate the glottal source from robust low-level speech features. In comparison to our proposed approach none of these methods discuss inverse filtering in a clustering/dictionary learning framework where the vocal tract parameters, along with the glottal excitations are jointly learnt from a large speech corpus.

III. Proposed Model for Glottal Inverse Filtering

A. Unsupervised Framework

Our framework follows from the well-known all-pole vocal tract model with P poles, where $x(n)$ is a speech sample at time n, $\{a_1, a_2, ..., a_P\}$ are the vocal tract parameters, $w(n)$ is the sparse glottal excitation derivative we wish to estimate, and $e(n)$ is white noise:

$$x(n) = \sum_{p=1}^{P} a_p x(n-p) + w(n) + e(n) \tag{2}$$

If we follow a vector representation of the speech samples for a window size of T samples, and assume the quasi-stationary nature of the vocal tract throughout the frame, then the speech production can be described by:

$$\mathbf{y} = \mathbf{Xa} + \mathbf{w} + \mathbf{e} \tag{3}$$

where $\mathbf{y}$ is the vector of samples $[x(P+1)\ x(P+2)\ x(P+3)...\ x(P+T)]$, $\mathbf{X}$ is the Toeplitz matrix constructed from the speech samples. In this expression, we have used the property that a convolution can be represented as multiplication with a Toeplitz matrix [9]. $\mathbf{w}$ and $\mathbf{e}$ are vectors representing the sparse glottal excitation derivative and white noise respectively. If we consider a collection of N frames, then for the $i-th$ frame:

$$\mathbf{y_i} = \mathbf{X_i a_i} + \mathbf{w_i} + \mathbf{e_i} \tag{4}$$

The white noise $\mathbf{e_i}$ for the i-th frame is modeled by a zero-mean Gaussian distribution with identity covariance matrix, thus we have:

$$\mathbf{y_i}|\mathbf{X_i}, \mathbf{a_i}, \mathbf{w_i} \sim \mathcal{N}(\mathbf{X_i a_i} + \mathbf{w_i}; \sigma^2\mathbf{I}) \tag{5}$$

Due to the spiky nature of the derivative $\mathbf{w_i}$ for the i-th frame, we impose a multi-variate Laplacian prior on $\mathbf{w_i}$ with location $\mathbf{0}$ and scale b.

$$\mathbf{w_i} \sim Laplace(\mathbf{0}, b) \tag{6}$$

We assume independence of $\mathbf{w_i}$, thus we have $\mathrm{P}(\mathbf{w_i}|\mathbf{X_i}, \mathbf{a_i}) = \mathrm{P}(\mathbf{w_i})$ and by the chain rule:

$$\mathrm{P}(\mathbf{y_i}, \mathbf{w_i}|\mathbf{X_i}, \mathbf{a_i}) = \mathrm{P}(\mathbf{y_i}|\mathbf{X_i}, \mathbf{a_i}, \mathbf{w_i}) \cdot \mathrm{P}(\mathbf{w_i}) \tag{7}$$

Expanding the probability distributions, and collapsing the parameters σ and b into a sparsity factor λ, we obtain the NLL (negative log-likelihood) for all frames as:

$$NLL = \sum_{i=1}^{N} \|\mathbf{y_i} - \mathbf{X_i a_i} - \mathbf{w_i}\|^2 + \lambda \|\mathbf{w_i}\|_1 \tag{8}$$

We also assume that the all-pole coefficients $\mathbf{a_i}$ are selected from a combination of K basis vocal tract filters $\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_K}$, so that we have $\mathbf{a_i} = \sum_{j=1}^{K} c_{ij}\mathbf{h_j} = \mathbf{Hc_i}$, where $\mathbf{H}$ is a matrix of basis filters. From a probabilistic interpretation of sparse coding [6], for the entire dataset of N frames we formulate the L_1 constrained loss as follows:

$$L = \sum_{i=1}^{N} \|\mathbf{y_i} - \mathbf{X_i H c_i} - \mathbf{w_i}\|^2 + \lambda \|\mathbf{w_i}\|_1 \tag{9}$$

where λ is a hyper-parameter controlling the amount of sparsity in the residual. For simplicity we have assumed a'winner-takes-all'configuration, where the vocal tract filter for each frame is contributed to by only one basis filter. This is enforced by a one-hot encoding scheme in $\mathbf{c_i}$ for the i-th frame. If we denote M_i as the cluster to which the i-th frame belongs, then $c_{ij} = 1$ when $j = M_i$, and zero otherwise. While this assumption greatly aids interpretability, we note that this model can be extended to a sparse combination of basis filters, where we can introduce an additional regularization term to control the sparsity in $\mathbf{c_i}$. Training the model corresponds to an optimization of the loss L over the unknown glottal excitation derivatives $\{\mathbf{w_i}\}$ and the dictionary $\mathbf{H}$, given the signal information in $\{\mathbf{y_i}, \mathbf{X_i}\}$ for a training set of N frames.

B. Parameter Learning

Under the 'winner-takes-all'assumption, the model formulation is similar to the K-means clustering problem, and can be learnt by an iterative EM (Expectation Maximization) style algorithm. The training algorithm is presented in Algorithm 1. We initialize training using random estimates of $\mathbf{H}$ and $\{\mathbf{w_i}\}$, and then update all parameters in turn for each iteration. The membership assignment corresponds to the E-step, and the estimation of $\mathbf{H}$ and $\{\mathbf{w_i}\}$ corresponds to the M-step. After the filterbank and basis memberships for each frame are estimated, the sparse residual $\mathbf{w_i}$ for the i-th frame can be estimated in the L_1 regularized LASSO framework [14] using an optimizer such as LARS (Least-Angle Regression) [15]. We continue iterating till loss function convergence. In practice we observe that 3-5 iterations are sufficient for convergence.

C. Estimation of glottal excitation for test datasets

Model testing is performed by keeping the learnt dictionary in $\mathbf{H}$ constant, and then iteratively updating the cluster memberships and the glottal excitation derivatives $\{\mathbf{w_i}\}$ for all frames of the testing dataset. This procedure is similar to model training as described in Table 1, with the exception of dictionary updates. Since $\mathbf{w_i}$ corresponds to the extracted glottal flow derivative of the i-th testing frame, it is necessary to compensate for effects of the lip radiation (which is generally modeled with a differentiator with a zero close to unity). Thus, we filter the estimated $\mathbf{w_i}$ for the i-th frame by a single-pole integrator with a transfer function $H(z) = \frac{1}{1-0.96z^{-1}}$ to obtain the glottal excitation signal.

D. Hyperparameter validation

The unsupervised framework has two hyperparameters: (1) Dictionary size K and (2) Sparsity factor λ characterizing the sparsity of the glottal excitation derivative. We employ a validation based approach to finding the optimal hyperparameters. Given a dataset, we split it into a speaker-independent training and validation set. For each K in the set {5,10,15,...,100} and sparsity λ in $\{10^{-1}, 10^{-2}, ..., 10^{-10}\}$, we train the model, and evaluate its performance on the validation set. The error metric used here is the log-spectral distortion between each frame in the validation set and its estimated spectral envelope [9]. From the validation experiments, we obtain the best hyperparameter set as $\{K = 80, \lambda = 10^{-7}\}$, which we shall use in all subsequent experiments.

Algorithm 1 Iterative Algorithm for Parameter Training

N: Number of frames; K: Size of dictionary
T: Frame duration in samples; P: Number of poles
$\mathbf{y}_i$: Signal vector for i-th frame
$\mathbf{X}_i$: Toeplitz data matrix for i-th frame
$\mathbf{H} : [\mathbf{h_1 h_2} ... \mathbf{h_K}]$ Dictionary
$\mathbf{W} : [\mathbf{w_1 w_2} ... \mathbf{w_N}]$ Excitation matrix
Initialize Dictionary: $\mathbf{H} \leftarrow rand(P,K)$
Frame memberships: $M_i \leftarrow None, i \in \{1,2,...,N\}$
Initialize Clusters: $C(j) \leftarrow \{\}, j \in \{1,2,...,K\}$
Initialize Excitation Matrix: $\mathbf{W} \leftarrow zeros(T,N)$
while loss not converged **do**
    1. Assign to Clusters
    for $i \in \{1,2,...,N\}$ **do**
        $M_i \leftarrow \arg\min_j \|\mathbf{y_i} - \mathbf{X_i h_j} - \mathbf{w_i}\|$
        $C(M_i) \leftarrow C(M_i) \cup \{i\}$
    end for
    2. Update Basis H
    for $j \in \{1,2,...,K\}$ **do**
        $\mathbf{h_j} \leftarrow (\sum_{i \in C(j)} \mathbf{X_i}^T \mathbf{X_i})^{-1} \sum_{i \in C(j)} \mathbf{X_i}^T (\mathbf{y_i} - \mathbf{w_i})$
    end for
    3. Update Excitation w
    for $i \in \{1,2,...,N\}$ **do**
        $\mathbf{w_i} \leftarrow \arg\min_w \|\mathbf{w} + \mathbf{X_i h_j} - \mathbf{y_i}\|^2 + \lambda \|\mathbf{w}\|_1$
    end for
end while

IV. Experimental Setup

In this section, we describe the datasets used in our experiments, along with the algorithms with which we have compared our approach. We have performed experiments not only on datasets of vowels, but also on phonetically balanced utterances spoken by multiple speakers. Since the vocal tract dictionaries are learnt from real-life speech corpuses, we have not tested our model on synthetic vowel datasets. We have also split the datasets in a speaker-independent manner, where the set of speakers for validation/testing are different from training set speakers. This also facilitates evaluation of the model's robustness to unknown speakers and speaking conditions.

A. Datasets

We have selected three datasets for conducting our experiments:
(1) *Finnish Vowels Dataset* : This dataset consists of recordings used in [16] spoken by 6 female and 5 male speakers aged between 18 and 48 years. Eight Finnish vowels were spoken across breathy, normal and tense voice qualities. Each participant repeated the same vowels thrice, producing 792 segments in the dataset. The training set consists of six speakers (three male, three female) and the validation set consists of four speakers (two male, two female).
(2) *CMU Arctic Dataset* [17] : We have selected 1132 utterances of the CMU Arctic dataset spoken by a single male speaker in US English. We have chosen this dataset for training purpose to compare the quality of dictionaries learnt on single speaker data, compared to multiple speaker data.
(3) *Cereproc Dataset* [18] : We have chosen a subset of the Cereproc dataset spoken by seven speakers in Received Pronunciation English. There are 315 utterances in the subset; where each speaker speaks the same set of 45 carefully designed phonetically balanced utterances. Data from five speakers (225 utterances) was chosen as the training set, and from two remaining speakers (95 utterances) as a validation set for investigating the generalization properties of the proposed model.

B. Models

We have compared the performance of our proposed model with the IAIF (Iterative Adaptive Inverse Filtering) algorithm proposed in [4]. For all algorithms, we have set the same all-pole model order of $P = 20$, and applied a Hanning window on each frame prior to analysis.

C. Methodology

We have designed experiments to address the primary research questions posed in this paper. To address question Q1, we have constructed an unsupervised approach for glottal inverse filtering in Section III. We further visualize the learnt dictionary (filterbank), and examine whether the basis filters can characterize vocal tract formants. We also compare the spectrum of a test voiced frame of speech with its corresponding LPC spectrum and the all-pole spectrum estimated from our model. To address research questions Q2 and Q3, thereby

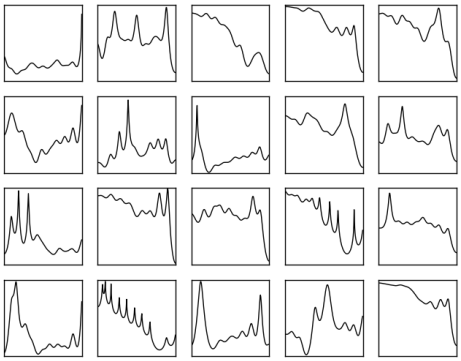


Fig. 1: Dictionary of basis vocal tract filters (power spectrum) learnt by the proposed model on Cereproc dataset. The maximum frequency displayed is 8 KHz for a sampling frequency of 16 KHz

determining whether the model learns meaningful glottal excitations and can generalize to unseen speakers, experiments comparing the glottal excitation waveforms with IAIF [4] are also performed. In these experiments we train on single and multi-speaker datasets (Cereproc train and CMU Arctic), and validate on the Finnish vowels and Cereproc validation datasets. Since the Finnish vowels dataset was sampled at 8 KHz, it was necessary to downsample any training corpus from 16 KHz when testing on the vowels dataset. It has been reported in the literature that phonation affects the glottal pulse shape and thus in our study, performance is evaluated across three categories of phonation: *breathy*, *modal* and *tense* voices. We are interested to examine if our proposed model generalizes better to certain voice quality categories.

V. Results and Discussions

A. All-pole filter dictionaries learnt

We performed training over the training set of speech data for five speakers from the Cereproc dataset with a dictionary size of $K = 20$ for visualization purposes. The learnt filters are presented in Fig. 1. The model is able to learn a compact set of filters, where sharp peaks in the spectra correspond to formants, and can thus approximate variations in the vocal tract across different speakers. It is to be noted that unlike the TVLP (Time Varying Linear Prediction) approach, where the filters are defined using fixed templates, we have obtained the dictionary in a data-driven manner for a large amount of data. Fig. 2 shows the periodogram of a test vowel frame /i/ from the CMU Arctic dataset, with the estimated spectral envelopes estimated by our proposed approach, and LPC. It is interesting to note that at higher frequencies, the spectral envelope is almost identical, probably due to a Hanning window being applied to the frame prior to the all-pole spectral estimation.

B. Glottal excitation waveforms

We consider three sets of training data for training the model - (1) Cereproc training data (five speakers) (2) CMU Arctic single speaker dataset and (3) Finnish Vowels dataset. Fig. 3 shows the speech waveforms corresponding to three vowel

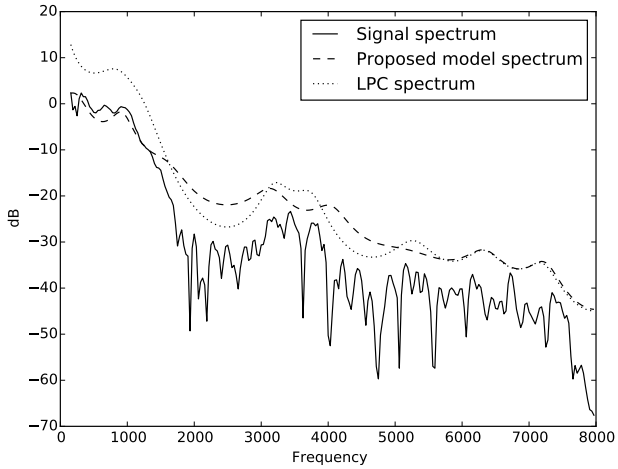


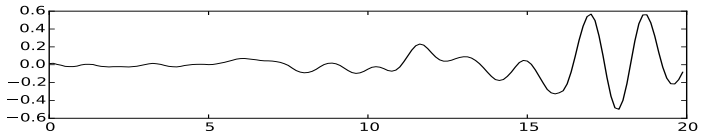
Fig. 2: Spectrum of a sample vowel /i/ from the CMU Arctic dataset, with envelopes estimated using LPC and our proposed method

frames from Finnish vowels and Cereproc validation dataset, along with the glottal flow estimated from our proposed model, and the IAIF approach. From an examination of the waveforms we observe that the glottal flow extracted in a data-driven unsupervised approach is comparable to IAIF, where information only from the local frame samples are utilized. It is also to be noted that we have enforced the sparsity constraint on the glottal flow derivative $\{\mathbf{w_i}\}$, so the glottal flow waveforms are not sparse themselves.

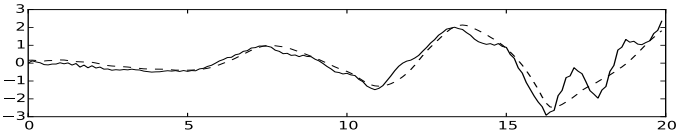
We further evaluate the quality of the extracted glottal source waveforms by comparing their similarity to the IAIF estimates in Table I. We choose the following metrics for measurement: (1) Log-spectral distortion (LSD) [9] and (2) Pearson's correlation coefficient (computed over each frame) and also analyze correlation across three phonation types: *breathy*, *modal* and *tense*. The mean, median and standard deviation of each metric are presented in the table for each dataset and phonation combination. We observe that there is a high correlation (and consequently low distortion) between the glottal excitations and the IAIF estimates, particularly when validated on the Finnish vowel dataset. Correlation is higher for breathy and modal voices, compared to tense voices. We hypothesize that this could be related to the degree of sparsity in the glottal flow derivative for different phonation types and the model has to be tuned accordingly. It is to be noted that our main focus is to develop an unsupervised data-driven framework for glottal inverse filtering, rather than outperforming existing state-of-the-art approaches.

VI. Conclusion

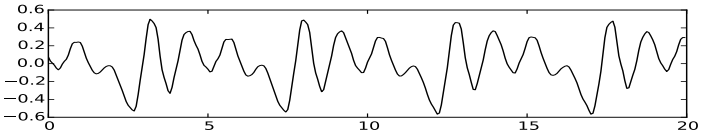
In this paper, we propose an unsupervised learning approach for glottal inverse filtering, in which the vocal tract filters and the excitation are jointly estimated in a data-driven fashion. We also compare the performance of our approach with the IAIF algorithm, and show that there is a low log-spectral distortion and high Pearson correlation between glottal source waveforms estimated by the two approaches, when validated



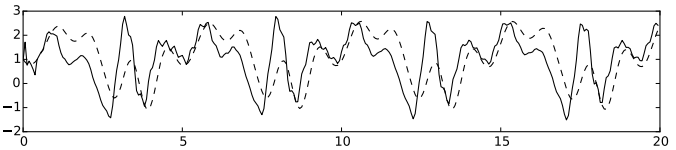
(a) Frame denoting start of vowel /a/ from Finnish vowel dataset



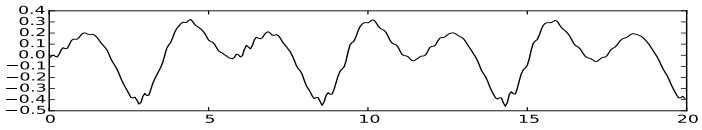
(b) Glottal excitation: proposed model (solid line) and by IAIF (dotted line)



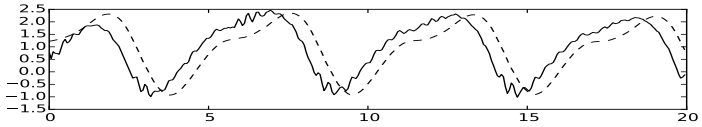
(c) Frame of vowel /a/ from Cereproc validation dataset



(d) Glottal excitation: proposed model (solid line) and by IAIF (dotted line)



(e) Frame of vowel /i/ from Cereproc validation dataset



(f) Glottal excitation: proposed model (solid line) and by IAIF (dotted line)

Fig. 3: Frames of vowel sounds from the Finnish vowel and Cereproc validation sets, with estimated glottal flow velocity waveforms from our proposed model and IAIF

on real-life speech databases. Validation experiments also show that the learnt dictionaries can generalize to unknown speakers. The proposed approach can be extended to neural networks for unsupervised learning, such as autoencoders, and a dictionary learning framework to learn glottal pulse descriptors, which could be discriminative for paralinguistic attributes such as emotion and voice quality.

ACKNOWLEDGMENT

The work depicted here is sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the U.S. Government and no official endorsement should be inferred. The authors wish to thank Prof. Paavo Alku and Dr. Matthew Aylett for providing the datasets used in this paper.

TABLE I: Correlation between IAIF and our proposed approach. Statistics are presented in the format mean, median (standard deviation)

Train	Val	Metric	Breathy	Modal	Tense
Finnish Train	Finnish Val	LSD(dB)	**9.39,8.77 (3.08)**	10.15,9.52 (3.22)	11.52,10.78 (3.58)
		Pearson	**0.71,0.74 (0.199)**	0.71,0.78 (0.197)	0.68,0.72 (0.188)
Cereproc Train	Finnish Val	LSD(dB)	9.51,9.27 (3.02)	**10.27,9.73 (3.09)**	11.18,10.53 (3.31)
		Pearson	0.56,0.66 (0.319)	**0.62,0.71 (0.304)**	0.59,0.67 (0.294)
Arctic Train	Finnish Val	LSD(dB)	**8.87,8.64 (2.36)**	9.40,9.94 (3.02)	10.75,10.14 (3.09)
		Pearson	**0.75,0.79 (0.19)**	0.71,0.75 (0.19)	0.65,0.69 (0.20)
Cereproc Train	Cereproc Val	LSD(dB)	**11.44,10.52 (4.27)**	11.85,10.94 (4.44)	12.14,11.26 (4.47)
		Pearson	**0.29,0.43 (0.38)**	0.26,0.38 (0.37)	0.29,0.38 (0.39)

REFERENCES

[1] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[2] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 21, no. 5, pp. 417–427, 1973.

[3] H. Fujisaki and M. Ljungqvist, "Estimation of voice source and vocal tract parameters based on arma analysis and a model for the glottal source waveform," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1987.*, vol. 12, pp. 637–640.

[4] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, pp. 109–118, 1992.

[5] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2528–2535.

[6] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[7] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (celp): High-quality speech at very low bit rates," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1985.*, pp. 937–940.

[8] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4650–4653.

[9] R. Giri and B. Rao, "Block sparse excitation based all-pole modeling of speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 3754–3758.

[10] A. Casamitjana, M. Sundin, P. Ghosh, and S. Chatterjee, "Bayesian learning for time-varying linear prediction of speech," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pp. 325–329.

[11] S. R. Chetupalli and T. V. Sreenivas, "Time varying linear prediction using sparsity constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6290–6293.

[12] J. Kane, S. Scherer, L.-P. Morency, and C. Gobl, "A comparative study of glottal open quotient estimation techniques," in *Proceedings of Interspeech 2013*. ISCA, 2013, pp. 1658–1662.

[13] M. Airaksinen, T. Raitio, and P. Alku, "Noise robust estimation of the voice source using a deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5137–5141.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[15] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[16] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proceedings of Interspeech 2007*.

[17] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[18] J. Kane, S. Scherer, M. Aylett, L.-P. Morency, and C. Gobl, "Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7982–7986.