Interspeech 2018
2-6 September 2018, Hyderabad

Time-regularized linear prediction for noise-robust extraction of the spectral envelope of speech

Manu Airaksinen, Lauri Juvela, Okko Räsänen, Paavo Alku

Aalto University, Finland

manu.airaksinen@aalto.fi, paavo.alku@aalto.fi

Abstract

Feature extraction of speech signals is typically performed in short-time frames by assuming that the signal is stationary within each frame. For the extraction of the spectral envelope of speech, which conveys the formant frequencies produced by the resonances of the slowly varying vocal tract, an often used frame length is within 20–30 ms. However, this kind of conventional frame-based spectral analysis is oblivious of the broader temporal context of the signal and is prone to degradation by, for example, environmental noise.

In this paper, we propose a new frame-based linear prediction (LP) analysis method that includes a regularization term that penalizes energy differences in consecutive frames of an all-pole spectral envelope model. This integrates the slowly varying nature of the vocal tract as a part of the analysis. Objective evaluations related to feature distortion and phonetic representational capability were performed by studying the properties of the mel-frequency cepstral coefficient (MFCC) representations computed from different spectral estimation methods under noisy conditions using the TIMIT database. The results show that the proposed time-regularized LP approach exhibits superior MFCC distortion behavior while simultaneously having the greatest average separability of different phoneme categories in comparison to the other methods.

Index Terms: speech analysis, linear prediction, robust features

1. Introduction

According to the classical source-filter theory by Fant [1], speech production can be modeled as a cascade of three processes, the glottal airflow excitation, the vocal tract and the lip radiation effect. By assuming that the vocal tract consists of concatenated lossless tubes, the vocal tract transfer function for non-nasalized sounds can be modeled with an all-pole transfer function. For low frequencies, the lip radiation effect is modeled as a time-derivative of the oral flow [2].

The glottal excitation can be simplified as a quasi-periodic signal with time-varying fundamental frequency and spectral tilt that carry information about individual speaker characteristics and paralinguistic cues [3]. For unvoiced speech, the glottal airflow can be modeled as white noise. Thus, the main apparatus in generation of linguistic information is the vocal tract, particularly the resonances of the tract, the formants [1]. The vocal tract consists of multiple *articulators* (e.g., tongue, soft palate, pharynx) that control its physical shape and thus the formant frequencies. Even though producing intelligible, continuously varying speech requires highly sophisticated and precise motor control of the articulators, the rate of changing the configuration of the vocal tract is limited by the inertial mass given by Newton's second law of motion. The traditional approach to account for the slow rate of change of the articulators (and thus formant contours of speech) is to assume speech to be stationary within short-time frames of approximately 20–30 ms and then process it in according time-frames with a frame skip of approximately 5–10 ms. This approach is default in speech feature extraction, mainly due to its simplicity, sufficient accuracy, and applicability for on-line processing. However, by performing speech feature extraction blindly frame-by-frame (e.g., with LP or MFCC computation), the obtained feature representation is oblivious to the context of the speech signal outside the frame. This makes frame-based feature extraction vulnerable to deterioration caused, for example, by background noise or the signal phase at the beginning of the frame [4]. Thus, *context-aware* speech feature extraction methods that take into account a broader "macro" context (e.g., 100 ms to few seconds) of speech have emerged.

Most prominent context-aware methods to compute autoregressive (AR) spectral envelope models of speech are time-varying linear prediction (TVLP) [5] and frequency-domain linear prediction (FDLP) [6]. In TVLP, the filter coefficient trajectories over a macro frame (either in direct form [5] or reflection coefficient form [7]) are fitted into a basis function of a pre-defined form (e.g., polynomial or trigonometric functions). In the FDLP approach, bandpass filtered time-trajectories of speech are modeled with an AR model, and these trajectories are used to compute frequency-domain parametric envelope estimates of speech at given time instants. Both TVLP and FDLP have been successfully applied for robust speech feature extraction with improved results over the baseline frame-based processing [8, 9], but their use is limited by the requirement of long macro frames that add algorithmic delay. Furthermore, the modeling of time-trajectories in both TVLP and FDLP is motivated by mathematical convenience, and not by arguing that, for example, the polynomial basis function in TVLP or the autoregressive model in FDLP would be the optimal model for the time-trajectory.

The present study proposes a novel regularization method in AR model computation in which the model optimization of a speech frame takes into account the filter coefficients of the previous frame, effectively resulting in a leaky integration process for the temporal dynamics of the obtained envelope spectrogram. This method, named time-regularized linear prediction (TRLP), yields smoothly evolving time-frequency contours akin to FDLP and TVLP, motivated by the slowness of mass movements of the articulators. Compared to FDLP and TVLP, the proposed TRLP does not add delay to the envelope modeling because long macro-frames are not needed. Results of this study indicate that TRLP shows superior performance compared to a set of known all-pole modeling techniques in spectral modeling of noisy speech.

10.21437/Interspeech.2018-1230

2. Time-regularized linear prediction

In this section, we introduce the TRLP method within the context of conventional LP. It should be noted that the use of the proposed method is applicable to all those forms of AR estimation methods (e.g., weighted LP [10], warped LP [11]) that have closed-form solutions or are obtained by the gradient-based optimization.

In the AR optimization, the conventional method is to minimize the residual loss. In LP, the squared error (i.e., the energy) of the residual is minimized:

$$\mathcal{L}_e = \frac{1}{2}\sum_n (s_n - \mathbf{a}^\mathsf{T}\mathbf{s_n})^2 \tag{1}$$

where $\mathcal{L}_e$ is the *residual loss*, s_n is the speech signal frame at sample n, the summation bounds n determine whether the covariance or autocorrelation criterion is used, $\mathbf{a} = [a_1, a_2, \ldots, a_p]^\mathsf{T}$ and $\mathbf{s_n} = [s_{n-1}, s_{n-2}, \ldots, s_{n-p}]^\mathsf{T}$. The gradient of the residual loss w.r.t. $\mathbf{a}$ is:

$$\nabla_\mathbf{a}\mathcal{L}_e = -\sum_n s_n\mathbf{s_n} + \left(\sum_n \mathbf{s_n}\mathbf{s_n}^\mathsf{T}\right)\mathbf{a}, \tag{2}$$

$$= -\mathbf{r} + \mathbf{R}\mathbf{a}, \tag{3}$$

where $\mathbf{r}^{(p\times 1)} = \sum_n s_n\mathbf{s_n}$ and $\mathbf{R}^{(p\times p)} = \sum_n \mathbf{s_n}\mathbf{s_n}^\mathsf{T}$. The conventional LP optimization is performed by setting Eq. 3 to zero and solving for $\mathbf{a}$. However, in the proposed method we add a regularization term $\mathcal{L}_{reg}$ into the loss function:

$$\mathcal{L}_{reg} = \frac{1}{2}\lambda_1(\mathbf{a} - \lambda_2\mathbf{a_{pr}})^\mathsf{T}(\mathbf{a} - \lambda_2\mathbf{a_{pr}}) \tag{4}$$

where $\mathbf{a_{pr}}$ denotes the filter coefficients of the previous frame, and λ_1 and λ_2 are regularization constants. $\lambda_1 \geq 0$ determines the overall effect of the regularization, and $\lambda_2 \in [0, 1]$ determines how much of the previous frame is used to weight the regularization. With $\lambda_2 = 0$, the regularization can be seen to simplify into conventional weight regularization (also known as Tikhonov regularization) [12, 13]. The gradient of $\mathcal{L}_{reg}$ w.r.t. $\mathbf{a}$ is:

$$\nabla_\mathbf{a}\mathcal{L}_{reg} = \lambda_1(\mathbf{a} - \lambda_2\mathbf{a_{pr}}) \tag{5}$$

$$= \lambda_1\mathbf{a} - \lambda_1\lambda_2\mathbf{a_{pr}} \tag{6}$$

Thus we can write the gradient of the regularized loss function and solve for $\mathbf{0}$:

$$\nabla_\mathbf{a}\mathcal{L} = \nabla_\mathbf{a}\mathcal{L}_e + \nabla_\mathbf{a}\mathcal{L}_{reg} = \mathbf{0} \tag{7}$$

$$-\mathbf{r} + \mathbf{R}\mathbf{a} + \lambda_1\mathbf{a} - \lambda_1\lambda_2\mathbf{a_{pr}} = \mathbf{0} \tag{8}$$

$$(\mathbf{R} + \lambda_1\mathbf{I})\mathbf{a} = \mathbf{r} + \lambda_1\lambda_2\mathbf{a_{pr}} \tag{9}$$

$$\mathbf{a} = (\mathbf{R} + \lambda_1\mathbf{I})^{-1}(\mathbf{r} + \lambda_1\lambda_2\mathbf{a_{pr}}) \tag{10}$$

It can bee seen from Eq. 10 that regularization affects only the diagonal elements of the matrix to be inverted. Thus, if the summation bound n in Eq. 1 with frame length N is set according to the autocorrelation ($n \in 0, 1, \ldots, N + p - 1$) or covariance criterion ($n \in 0, 1, \ldots, N - 1$), the resulting $p \times p$ matrix $\mathbf{R}$ will have an efficiently invertible Toeplitz structure (which is also symmetric for the autocorrelation criterion) [2] that is transferred also to the regularized matrix.

2.1. Choice of regularization terms

A factor to account for in the solution is the effect of the frame energy (which affects $\mathcal{L}_e$) relative to $\mathcal{L}_{reg}$, which are mainly controlled by the regularization constant λ_1. The solution can be made energy-invariant by normalizing the components $\mathbf{r}$ and $\mathbf{R}$ with frame energy $r_0 = \sum_n s_n^2$:

$$\mathbf{r}' = \mathbf{r}/r_0 \tag{11}$$

$$\mathbf{R}' = \mathbf{R}/r_0. \tag{12}$$

However, energy-invariance does not always lead to best performance: For example, for low-energy frames (where the ratio of speech energy to the noise floor is low) it might be beneficial to increase the effect of $\mathcal{L}_{reg}$ compared to $\mathcal{L}_e$ with an adaptive λ_1 and/or λ_2. This investigation is, however, outside the scope of the current study, and we will limit the experiments to the energy-invariant solution with constant regularization terms. Based on informal experiments, a reasonable range for the terms is $\lambda_1 \in [0.2, 2.0]$ and $\lambda_2 \in [0.9, 0.99]$. For the experiments presented in this study, we use the values $\lambda_1 = 1.0$ and $\lambda_2 = 0.9$.

3. Experiments

Robust AR methods have been successfully applied as a preprocessing step in computing mel-frequency cepstral coefficients (MFCCs) within front-ends of automatic speech recognition (ASR) [14] and speaker recognition systems [9, 8]: Instead of computing the mel-filter bank energies directly from the Fourier magnitude spectrum, the AR method is used to obtain a noise-robust, parametric estimate of the spectral envelope, from which the filter bank energies are computed. The MFCC representation of the spectral envelope is commonly preferred over alternative representation forms such as mel-filter bank energies or line spectral frequencies [15] because of the statistically convenient properties of the cepstral coefficients [16]. Keeping this in mind, we chose to perform our experiments with the MFCC representations obtained with the various methods. Our first experiment (reported in section 3.3) studies the MFCC distortion of obtained feature vectors under various noisy conditions, and in our second experiment (reported in section 3.4) we look at the MFCC vectors' distribution distances between phoneme categories under noisy conditions. In all of our experiments, we utilize the complete test set from the TIMIT corpus [17] which consists of 1,680 phonetically balanced utterances from 168 different speakers with varying American English dialects.

3.1. Noise conditions for test samples

Both of our experiments utilize the same noise conditions: The clean TIMIT test set utterances were artificially corrupted with three different noise types (speech-like, babble, and pink) with a global signal-to-noise ratio (SNR) varying from -5 dB to 20 dB with 5-dB increments. The speech-like noise was generated as spectrally shaped white noise, where the spectral envelope of the noise was set as the long-time average magnitude spectrum computed over the TIMIT train set. The babble noise was obtained from the NOISEX database [18]. All of the noise-corrupted samples were pre-generated for the tests to ensure identical noise conditions.

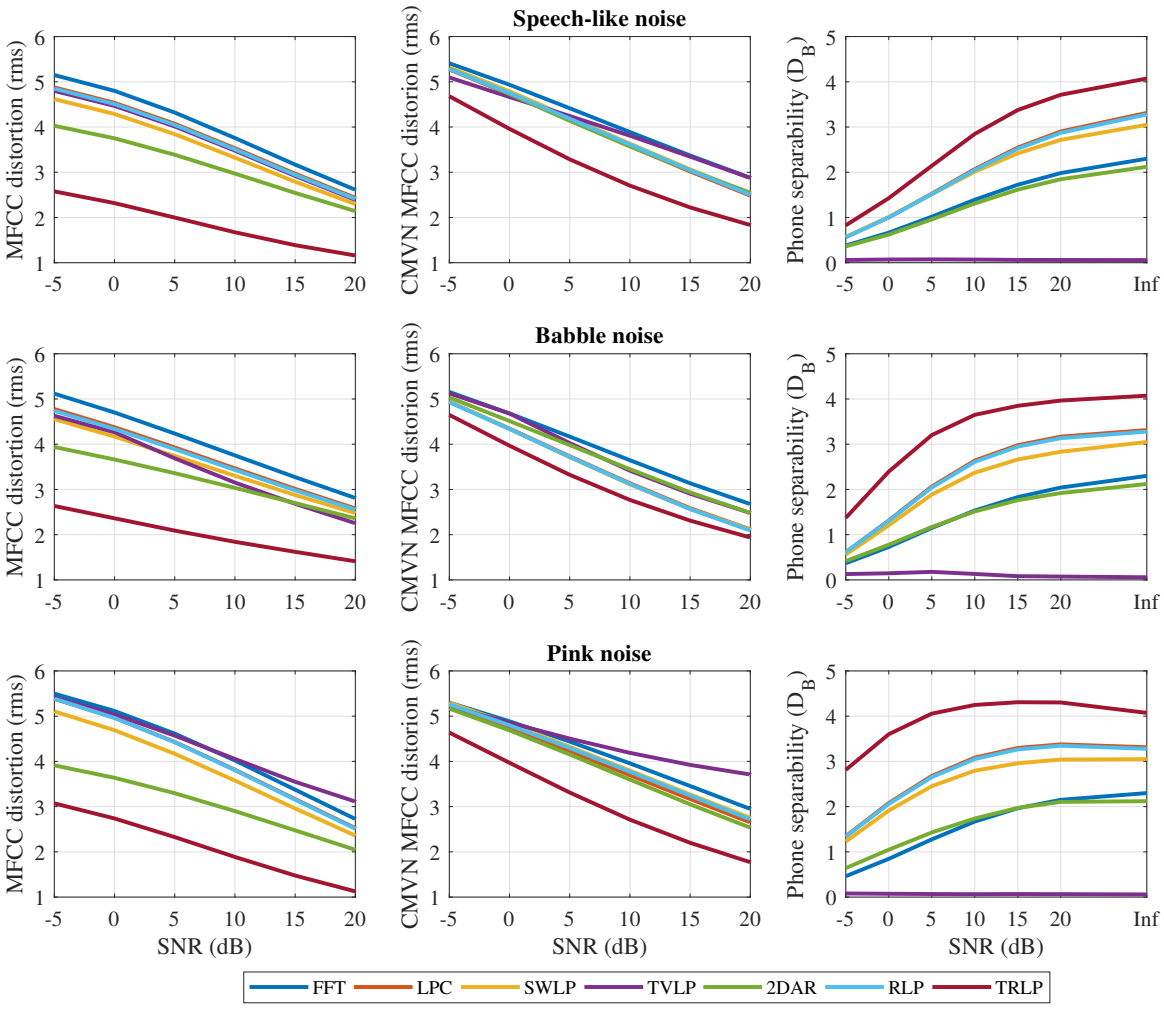

Figure 1: *Results of the MFCC distortion test. Rows correspond to different noise types. MFCC distortion is reported for the direct form (left column) and with CMVN (middle column). The Bhattacharyya distances D_B for phoneme category separability are shown in the right column.*

3.2. Reference methods and MFCC extraction

For the experiments, we chose a number of well-known AR-model -based reference methods, from which we compute the MFCCs. The reference methods include: FFT-based MFCCs, conventional LP [2], time-varying LP (TVLP) [7], FDLP (2DAR) [9], stabilised weighted LP (SWLP) [19], and regularized LP (RLP) [13]. Where applicable, the analysis parameters were set as follows: sampling rate $f_s = 16000$ Hz, window length 25 ms, Hamming window function, frame skip 10 ms, AR model order $p = 20$, and FFT length of 1024. In TVLP, a reflection coefficient-based formulation with 3rd order polynomial basis functions was utilized, with a macro-frame length of 200 ms with an overlap of 20 ms between macro-frames. In 2DAR, 1-second macro frames with 100 ms overlap were utilized, and the FDLP model order was set to 120. For SWLP, the order for short-term energy function was set to 20, and in RLP we utilized a constant $\lambda = 0.04 \cdot 10^{-3}$. After the acquisition of the AR filter polynomials, we compute the power spectrum of the all-pole filters with FFT, zero-padded to the desired FFT length, and ensure AR filter stability by adding a small constant of $\epsilon = 10^{-12}$ to each inverse filter magnitude spectrum before computing the synthesis filter.

The MFCC computation utilized a triangular HTK-style mel-filter bank with constant band-wise energies. Mel-band energies were computed using 24 channels, followed by computing the log and the discrete cosine transform (DCT). From the DCT we utilized the coefficients $c_1 - c_{19}$, thus omitting frame energy (contained by c_0).

3.3. MFCC distortion test

MFCC distortion is defined as the root-mean squared error of the noise corrupted frame's MFCC vector to the clean sample's MFCC vector:

$$D_{\text{MFCC}} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{M}\sum_{m=1}^{M}(c_{n,m} - \hat{c}_{n,m})^2}, \qquad (13)$$

where $c_{n,m}$ is the mth cepstral coefficient of the nth clean MFCC vector, $\hat{c}_{n,m}$ is its noise-corrupted version, N is the total number of MFCC vectors and M is the number of MFCC

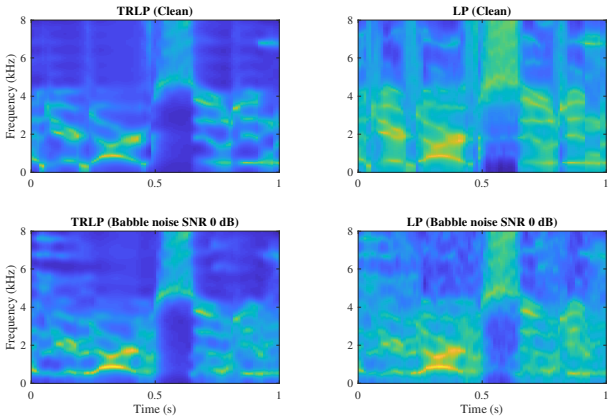

Figure 2: *Frame-wise energy-normalized envelope spectrograms for an one second segment of speech obtained with the proposed TRLP method (left) and conventional LP (right) in clean (top) and noisy (bottom) conditions.*

coefficients in each vector. As an alternative to straightforward computation of Eq. 13, cepstral mean and variance normalization (CMVN) [20] can be performed on the MFCC vectors to better account for a balanced distance measure across all feature dimensions. In our experiments, we compute the direct MFCC distortion as well as distortion after utterance-level CMVN.

3.4. Phoneme separability test

Related to the MFCC distortion measure, it is worth noting that even though it is desirable that a MFCC vector computed from a noisy frame would be as close as possible to the corresponding vector computed from the clean frame, a small distortion value does not always mean that the underlying spectral estimation method is robust. For example, one can easily envision a pathological case where a method produces completely flat spectral representations for all frames, and thus achieves zero distortion, while failing completely in the modeling of the underlying signal spectrum. For this reason, we have studied also the phoneme category distribution distances: An ideal spectral estimation method would maintain MFCC vectors corresponding to same phoneme close to each other while maximizing the distance to other phoneme classes, while simultaneously minimizing the MFCC distortion to maintain consistency across different SNRs.

In our tests, we used the standard reduced TIMIT phoneme set of 39 unique phonemes [21]. To obtain the phoneme-specific distributions, each feature vector was assigned to the phoneme category annotated for the center of the frame, and then the within-category distances between all frames were compared against the distances to frames from other phoneme categories. Since the MFCCs computed from speech approximately follow a multivariate Gaussian distribution [16], we utilized the Bhattacharyya distance D_B for multivariate Gaussian distributions [22] as the distance metric between two phoneme categories:

$$D_B = \frac{1}{8}(\mu_1-\mu_2)^\intercal \mathbf{\Sigma}^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\left(\frac{\det \mathbf{\Sigma}}{\sqrt{\det \mathbf{\Sigma}_1 \det \mathbf{\Sigma}_2}}\right), \quad (14)$$

where μ_i and $\mathbf{\Sigma}_i$ are the means and covariances of the distributions, and $\mathbf{\Sigma} = \frac{\mathbf{\Sigma_1}+\mathbf{\Sigma_2}}{2}$. Full covariance matrices were used in the computations.All possible pair-wise Bhattacharyya distances between unique phoneme classes were then computed for each compared method (a total of 741 distances) and averaged to obtain the final test metric. This was done separately for each noise condition.

4. Results

A representative example of obtained spectrograms is shown in Figure 2 for TRLP and conventional LP under clean and noisy conditions. It can be seen that TRLP traces the formants with smooth and continuous trajectories.

4.1. MFCC distortion test

The MFCC distortion results are presented in the first and second columns of Figure 1. We can see that the utilization of AR-based envelope estimation as a pre-processing step manages to decrease MFCC distortion over the FFT-baseline for all noise types. In both cases (direct and CMVN distortion) TRLP manages to have the smallest distortion, but we can see that it has the relatively greatest drop in performance when CMVN is applied.

4.2. Phoneme separability test

The Bhattacharyya distances (phoneme separability) averaged across all phoneme pairs are presented in the third column of Figure 1. Here, again, the use of AR-based envelope estimation can be seen to be beneficial over the FFT baseline, except in the case of TVLP, where the method with the used parameters fails to produce well separable distributions for the phoneme categories. Remarkably, the TRLP method achieves the highest separability of the phoneme categories while simultaneously having the smallest MFCC distortion for all noise levels and types. This confirms that the approach is not simply minimizing the cepstral distortion at the cost of representational capability.

5. Discussion

This study presents a time-regularized linear predictive method that penalizes the rate of change in the model coefficients between the current and previous frame. The method, called TRLP, was evaluated in terms of its feature distortion and phonetic representational capability under noisy conditions with MFCCs computed from the obtained spectral model. The results show that TRLP exhibits superior MFCC distortion behavior while simultaneously having the greatest average separability of phoneme categories in comparison to reference methods. This suggests that TRLP could be used as the spectral envelope model in a wide range of applications where noise robustness without compromised representational accuracy is required. The present study reports our first experiments on TRLP using general objective metrics that demonstrate the promise of the approach. However, further experiments with different speech processing applications are required to fully understand the applicability of the TRLP method.

6. Acknowledgements

The research leading to these results has received funding from the Academy of Finland (project no. 256961, 312490).

7. References

[1] G. Fant, *Acoustic Theory of Speech Production*. Mouton & Co., 1960.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[3] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1, pp. 189–212, 2003.

[4] D. O'Saughnessy, *Speech Communications – Human and Machine*. IEEE Press, 2000.

[5] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Processing*, vol. 5, no. 3, pp. 267–285, 1983.

[6] S. Ganapathy, S. H. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1285–1295, 2014.

[7] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 4, pp. 899–911, 1983.

[8] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, "Time-varying autoregressions for speaker verification in reverberant conditions," in *Proc. Interspeech*, 2017, pp. 1512–1516.

[9] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-D autoregressive models for speaker recognition," in *IEEE Speaker Odyssey*, 2012.

[10] C. Ma, Y. Kamp, and L. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 2, pp. 69–81, 1993.

[11] H. W. Strube, "Linear prediction on a warped frequency scale," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1071–1076, 1980.

[12] A. Tikhonov and V. Arsenin, *Solutions of ill-posed problems*, ser. Scripta series in mathematics. Winston, 1977.

[13] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.

[14] J. Pohjalainen, C. Magi, and P. Alku, "Enhancing noise robustness in automatic speech recognition using stabilized weighted linear prediction (SWLP)," in *Proc. Interspeech*, 2008.

[15] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[16] M. Airaksinen, *Analysis/Synthesis Comparison of Vocoders Utilized in Statistical Parametric Speech Synthesis*. Master's Thesis: Aalto University, 2012.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," in *LDC93S1 [Web Download]. Linguistic Data Consortium*, 1993.

[18] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[20] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133–147, 1998.

[21] M. Antal, "Speaker independent phoneme classification in continuous speech," *Studia Universitatis Babeş-Bolyai. Informatica*, vol. 49, no. 2, 2004.

[22] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967.