Mixture Linear Prediction in Speaker Verification Under Vocal Effort Mismatch

Jouni Pohjalainen, Cemal Hanilçi, Tomi Kinnunen, and Paavo Alku, *Senior Member, IEEE*

***Abstract*—This paper describes an approach to robust signal analysis using iterative parameter re-estimation of a mixture autoregressive (AR) model. The model's focus can be adjusted by initialization of the target and non-target states. The variant examined in this study uses an i.i.d. mixture AR model and is designed to tackle the spectral biasing effect caused by the voice excitation in speech signals with variable fundamental frequency. In our speaker verification experiments, this method performed competitively against standard spectrum analysis techniques in non-mismatch conditions and showed significant improvements in vocal effort mismatch conditions.**

***Index Terms*—Speech feature extraction, spectrum analysis, robust acoustic features, speaker recognition.**

I. Introduction

SPECTRUM analysis is essential in most signal processing applications. Methods such as the fast Fourier transform (FFT) and linear prediction (LP) perform well under ideal conditions but their performance typically degrades in the presence of distortions. In audio signal processing, the distortions include background noise (additive) and channel distortion (convolutive). In *speech* signal processing, spectral information is further affected by many speaker-related effects, such as speaking style, vocal effort, and fundamental frequency (F0). For instance, the formant locations produced by LP for high-pitch speech are biased towards the nearest F0 harmonics [1].

Distortion and speaker-related variability are particularly detrimental in speech applications that use machine learning, such as automatic speech and speaker recognition, in which short-time spectra are parametrized using statistical models. If the conditions of the model training differ from their actual usage in the test phase, the *mismatch* in the spectral feature statistics between the training and test might severely degrade the performance. *Multi-condition* training is often found to improve the performance, even in adverse conditions, but the use of this approach is limited by the difficulty and cost of collecting sufficient training data in order to cover all relevant usage conditions. Therefore, in order to improve signal analysis performance in mismatched and variable conditions, it is necessary to study spectrum analysis methods that are inherently robust with respect to the sources of variability and distortion present in real-world signals. This topic is addressed in the present study by first introducing a general principle for stochastic linear predictive analysis based on a mixture autoregressive model. We demonstrate that the method leads to iteratively generated temporal weighting of the signal information based on initial autoregression templates—a target template to look for and a non-target template to avoid.

The application being studied is text-independent speaker verification under vocal effort variability and mismatch. F0 increase is known to be one of the main acoustical effects when vocal effort is raised from soft to loud [2], [3]. With the advent of real machine learning applications for speech signals, tackling vocal effort mismatch is becoming increasingly important [4], [5]. Previously, feature-level solutions have been proposed to compensate for the effects of vocal effort other than F0, such as spectral tilt [6], [7]. However, previous studies on F0 robustness have concentrated primarily on formant estimation (e.g., [1], [8], [9]). Given the recent success of time-weighted linear prediction in robust feature extraction under additive noise (e.g., [10], [11]), it is justified to study the proposed stochastic, time-weighted linear predictive modeling approach in feature extraction by customizing it to produce spectra less affected by F0 variation than standard methods.

II. Linear Predictive Spectrum Estimation

A. Linear Prediction Weighted in Time

Linear predictive modeling assumes that the signal s_n follows a zero-mean autoregressive (AR) process $s_n = \sum_{k=1}^{p} a_k s_{n-k} + G u_n$ of order p, which in the z domain corresponds to an all-pole filter $H(z) = G/(1 - \sum_{k=1}^{p} a_k z^{-k})$. Here, u_n is the excitation signal and G is its optional gain [12]. Linear prediction (LP) solves the predictor coefficients a_k by minimizing the prediction error energy $\sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2$, where each prediction is a linear combination of the a_k and the previous samples. In this work, sums over n follow the autocorrelation method [12]. Weighted linear prediction (WLP) generalizes LP by instead minimizing a time-weighted energy $E_W = \sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2 W_n$ [9], which emphasizes "reliable" signal segments and de-emphasizes others; LP follows as a special case by making the weighting function W_n a constant. Typically, the short-time energy (STE) of recent samples, $W_n = \sum_{i=1}^{p} s_{n-i}^2$, is chosen for weighting in order to emphasize high-energy segments of the analysis frame that, with stationary background noise, have a high local signal-to-noise ratio. The coefficients are solved by setting $\partial E_{\mathrm{W}} / \partial a_j = 0$, leading to the normal equations $\sum_{k=1}^{p} a_k \sum_n W_n s_{n-k} s_{n-j} = \sum_n W_n s_n s_{n-j}$, $1 \leq j \leq p$.

J. Pohjalainen and P. Alku are with the Department of Signal Processing and Acoustics, Aalto University, Espoo, 00076 Finland (jpohjala@acoustics.hut.fi, paavo.alku@aalto.fi). C. Hanilçi is with the Department of Electrical-Electronic Engineering, Bursa Technical University, Bursa, 16190 Turkey (cemal.hanilci@btu.edu.tr). T. Kinnunen is with the School of Computing, University of Eastern Finland, Joensuu, 80101 Finland (tomi.kinnunen@uef.fi.)

This work has been supported by Academy of Finland (256961, 253120) and the EC FP7 project Simple4All (287678).

STE weighting emphasizes, within the pitch period, the beginning of the glottal closed phase, where formants are prominent (see Fig. 1). However, each closed phase begins with transient, high-amplitude samples at the glottal closure instant (GCI) as the main acoustical excitation of the vocal tract is generated. These transient samples at GCIs do not contain formant information. As F0 increases, they cover a larger proportion of the frame. In the frequency domain, this is marked by the spectral harmonic structure becoming sparse, with F0 harmonics biasing the modeling of formants using conventional spectrum analysis techniques. STE weighting does not specifically downweight GCIs, resulting in F0 bias persisting also in such WLP models. It was recently demonstrated in [1] that F0 bias in formant estimates can be reduced by, instead, designing a weighting function W_n that downweights the squared residual around the GCIs. However, while algorithms have been proposed to explicitly estimate GCIs, these epochs are difficult to estimate reliably. In [1], the GCI information was obtained in an oracle-like manner by using synthetic vowels as test material. In the next sections, we propose an automatic method to *simultaneously* estimate an appropriate weighting function and the all-pole coefficients.

B. Mixture Linear Prediction

1) The General Mixture Autoregressive Model: The signal s_n, $n \geq 0$, can be modeled as a mixture of J autoregressive processes with conditional density $f(s_n|s_{n-1},\ldots,s_0,\lambda) = \sum_{i=1}^{J} \pi_{n,i} \frac{1}{\sigma_i} \phi\left(\frac{u_{n,i}}{\sigma_i}\right)$, where λ is the model's parameter set and $\phi(\cdot)$ is the standard normal density function; $\pi_{n,i} = P(q_n = i|s_{n-1},\ldots,s_0,\lambda)$, $1 \leq i \leq J$, is the *prior* distribution of a hidden state variable $q_n \in \{1,\ldots,J\}$ that determines which one of the J AR processes,

$$s_n = a_{0,i} + \sum_{k=1}^{p} a_{k,i} s_{n-k} + u_{n,i}, \qquad 1 \leq i \leq J, \tag{1}$$

generates sample s_n. The $a_{0,i}$ are intercept (constant) terms and the $u_{n,i} \sim \mathcal{N}(0, \sigma_i^2)$ are Gaussian white noise excitations.

Two main approaches to modeling the latent state process q_n exist and have been previously studied in time series analysis and econometrics: q_n can be considered i.i.d. and modeled using *component weights* as $\pi_{n,i} = P_i$ $\forall n$, leading to an *i.i.d. mixture* AR model [13], or it can be assumed to follow a first-order Markov process, leading to a linear predictive hidden Markov or *Markov-switching* AR model [14], [15], [16], [17].

In speech processing, mixture AR models appear not to have been previously applied to frame-level spectrum analysis, but similar models have been used for parametrizing utterances in recognition applications. Some of them work on the feature vector level, [18], [19], while others, [20], [21], work on the signal level but apply the AR dynamics in separate frames. Some Markov-switching recognition models, [22], [23], are similar to the current signal model in that they consider each *sample* and its associated hidden state q_n. The current method differs from previous studies both by applying the signal model to frame-level spectrum analysis and by being directed towards finding an AR model of a target type as opposed to a non-target type. Before parameter estimation using the iterative expectation-maximization (EM) principle [24], one of the states is designated as *target*, the other one(s) as *non-target* and the AR parameters of these states are initialized with simplified, characteristic descriptions of desired and undesired signal qualities, respectively. Because EM increases the model likelihood with each iteration, it will converge towards a local likelihood maximum whose location on the parameter hypersurface is determined by the initial parameter values.

2) Dynamics of the Hidden State Process: In this study, we investigate *Gaussian mixture linear prediction* (GMLP) [25], a simple implementation of the targeted mixture principle. The probability law governing q_n is assumed to be i.i.d. and parametrized with component weights $\pi_{n,i} = P_i$, similarly to Gaussian mixture models (GMMs) [26]. The GMLP signal model is thus specified by the set of parameters $\lambda_{\text{GMLP}} = (P_1,\ldots,P_J,a_{0,1},a_{1,1},\ldots,a_{p,1},a_{0,2},\ldots,a_{p,J},\sigma_1^2,\ldots,\sigma_J^2)$, iteratively re-estimated by applying the EM principle [24]:

1) In the E step, estimate the excitations $u_{n,i}$ as prediction residuals $e_{n,i} = s_n - a_{0,i} - \sum_{k=1}^{p} a_{k,i} s_{n-k}$. Then, determine the hidden state *posterior* probabilities $\gamma_{n,i} = P(q_n = i|s_n,\ldots,s_{n-p},\lambda_{\text{GMLP}}) = \max\left(0.01, \frac{P_i(1/\sqrt{2\pi\sigma_i^2})\exp(-e_{n,i}^2/(2\sigma_i^2))}{\sum_j^J P_j(1/\sqrt{2\pi\sigma_j^2})\exp(-e_{n,j}^2/(2\sigma_j^2))}\right)$ and renormalize so that $\sum_i \gamma_{n,i} = 1$ $\forall n$ (a lower limit of 0.01 is imposed to avoid occurrence of unused states).
2) In the M step, re-estimate the component weights as $P_i = \frac{\sum_n \gamma_{n,i}}{\sum_n 1}$ and the variances as $\sigma_i^2 = \frac{\sum_n \gamma_{n,i} e_{n,i}^2}{\sum_n \gamma_{n,i}}$. For $a_{k,i}$, define $x_{n,0} = 1$ (for the intercept) and $x_{n,k} = s_{n-k}$, $k \geq 1$, and solve the normal equations $\sum_{k=0}^{p} a_{k,i} \sum_n \gamma_{n,i} x_{n,k} x_{n,j} = \sum_n \gamma_{n,i} s_n x_{n,j}$, $0 \leq j \leq p$. Barring the intercept term, the latter equations are equivalent to standard WLP (Section II-A) weighted by corresponding state posterior probabilities ($W_n = \gamma_{n,i}$).

Equivalent formulas are given in [13]. Notably, setting the AR order $p = 0$ makes the intercepts behave like Gaussian means and leads to conventional GMM re-estimation formulas [26].

In each iteration of GMLP, the time complexity of the E step (determination of $e_{n,i}$ and $\gamma_{n,i}$, $1 \leq i \leq J$) is $O(JNp)$, where N is the number of samples within the analysis frame. In the M step, re-estimation of P_i and σ_i^2 is $O(JN)$. Apart from the factor J (the number of states), the above operations are of the same order as windowing and correlation in both GMLP and classical linear predictive methods [27]. The remaining computation in one iteration is due to solving the J groups of weighted normal equations for the AR coefficients $a_{k,i}$, which is $O(Jp^3)$ by using the Cholesky decomposition, also used with the covariance method of LP [27]. The computation load of GMLP relative to standard methods thus depends linearly on the number of iterations and states.

In *Markov-switching linear prediction* (MSLP), the component weights P_i are replaced by two sets of parameters: state transition probabilities $p_{i,j}$ and initial state probabilities ρ_i [16]. Again, parameter estimation is iterative and based on EM. However, it needs to use the computationally more expensive forward–backward algorithm [26], or another similar approach [17], to compute the probabilities required for re-estimating the model parameters. In preliminary tests on different systems, this noticeably increased the feature computation time, but did not improve the verification performance. Thus, only GMLP is evaluated in this study.

3) Role of the Constant Terms: In conventional LP, intercept terms are not used. The intercept is zero for a zero-mean AR process [16] and can thus be omitted. For speech, the assumption of a zero mean approximately holds true when using analysis frames large enough to cover more than one pitch period, since speech does not contain important frequencies below F0. In mixture linear prediction, however, the inclusion of the intercept term (Eq. 1), even if initialized with zero for each state, allows the AR models the freedom to focus on subsets of the analysis frame without implicitly assuming their samples to add to zero. Moreover, it is possible that the frame would have a non-zero mean due to low-frequency distortion components. Despite the inclusion of the intercept terms in the EM iteration, this term is not included in the final target all-pole model which is chosen as $H(z) = 1/(1 - \sum_{k=1}^{p} a_{k,1} z^{-k})$.

4) Application to Speech with Variable F0: While different initializations of the mixture model lead to different target models, in this study, we concentrate on downweighting GCIs [1]. With $J = 2$, we initialize the target model (state 1) with $a_{k,1} = 0$, $k \neq 1$, and $a_{1,1} = 0.97$. It thus corresponds to the single-pole filter $1/(1 - 0.97z^{-1})$, the inverse of the typical pre-emphasis filter $1 - 0.97z^{-1}$ used to compensate for the spectral tilt of voiced speech. This can be viewed as a rough approximation of the characteristic low-pass spectrum envelope of voiced speech. For the non-target state 2, the lagged AR parameters are initialized with $a_{k,2} = 0$, $k \geq 1$, and the intercept with $a_{0,2} = \max(s_n)$. This filter has a flat spectrum, like the spectrum envelope of an impulse train, and a focus on large signal values typical at GCIs. As Fig. 1 shows, during the EM iteration the target state 1 gravitates towards signal segments that have low-pass spectral characteristics and smaller amplitudes, while state 2 concentrates on GCIs, collecting their effects and preventing the harmonics of F0 from biasing the target model spectrum. For this initialization, it has been found beneficial to perform one preliminary iteration of EM where only P_i and σ_i^2 are updated (from the initial values of $P_i = 0.5$ and $\sigma_i^2 = 0.01$, $1 \leq i \leq 2$).

III. Experiments

A. Experiment Setup

The experiments are carried out on core tasks of the 2010 NIST SRE corpus involving conversational telephone speech sampled at 8 kHz. We examine three vocal effort conditions:

- **Det 5: Normal vocal effort** in both training and test, test data containing 708 target and 29655 impostor trials.
- **Det 6: Normal vocal effort** in training and **high vocal effort** in test (361 target and 28311 impostor trials).
- **Det 8: Normal vocal effort** in training and **low vocal effort** in test (289 target and 28306 impostor trials).

A GMM-supervector/support vector machine (SVM) subsystem [28] with channel compensation by nuisance attribute projection (NAP) [29] is used as a quick-to-train classifier to experiment with the number of GMLP iterations and to add to the generality of the tests. Gender-dependent UBMs (universal background models) with 512 Gaussians are trained using the NIST SRE05, SRE06, and Switchboard corpora. Negative examples (background speakers) to train speaker-dependent SVMs are selected from the SRE03 and SRE04 corpora (395 male and 577 female speech files). NAP matrices are trained using 2020 male and 2017 female utterances from SRE06. In adapting the mean vectors, relevance factor $r = 8$ is used.

In the i-vector [30] system, we use gender-dependent UBMs with 1024 Gaussians, trained on the same material as those for the GMM-SVM system. Gender-dependent T-matrices (or i-vector extractors) are trained with 5 EM iterations using the SRE04, SRE05, SRE06, Fisher, and Switchboard corpora (19084 male and 24237 female utterances). 600-dimensional, whitened and length-normalized i-vectors are extracted for each utterance and compared using a probabilistic LDA (PLDA) back-end with a 200-dimensional speaker subspace.

Mel-frequency cepstral coefficients (MFCCs) for the classifiers are computed in Hamming-windowed frames with length 30 ms and overlap 15 ms. Spectra given by FFT, LP ($p = 20$), WLP ($p = 20$), or the described variant of GMLP ($p = 20$) are processed as follows: 1) square the magnitude spectrum, 2) multiply it by 27 triangular filters spaced evenly on the mel scale, 3) take the logarithm of the filterbank output energies, and 4) apply discrete cosine transform to obtain 18 MFCCs without the zeroth coefficient. Next, the MFCCs are RASTA filtered across frames and $\Delta/\Delta\Delta$ features are appended to the feature vectors [11]. Finally, utterance-level cepstral mean and variance normalization and voice activity detection based on frame energies are applied to the feature vector sequence.

B. Results

The systems are evaluated at two operating points (detection thresholds) determined by the miss and false alarm rates p_{miss} and p_{fa}. Tables I and II show equal error rates (EER; $p_{\text{miss}} = p_{\text{fa}}$) and minimum decision cost function (MinDCF; minimal $0.1p_{\text{miss}} + 0.99p_{\text{fa}}$) values, respectively. Auxiliary GMM-SVM tests suggest that more GMLP iterations may be optimal for mismatched female speech than otherwise, likely due to greater F0 bias. Statistical analyses of differences in $0.5p_{\text{miss}} + 0.5p_{\text{fa}}$ at 95 % confidence [31], for the i-vector system at both thresholds separately for male and female speech (12 cases), show GMLP to significantly outperform FFT, LP and WLP on female speech with vocal effort mismatch (DET 6 and DET 8), while FFT outperforms the other methods in the DET 8 male case. GMLP is in the best performing group (with no significant differences) for DET 5 male at both thresholds and for DET 5 female and DET 6 male at the EER and MinDCF thresholds, respectively. These results suggest that this variant of GMLP performs competitively in matched vocal effort conditions – still containing F0 variation and potential mismatch – and especially support its use for improving system robustness against mismatch caused by raised vocal effort. Such conditions can easily occur in noisy real-world environments due to, e.g., the Lombard reflex [32].

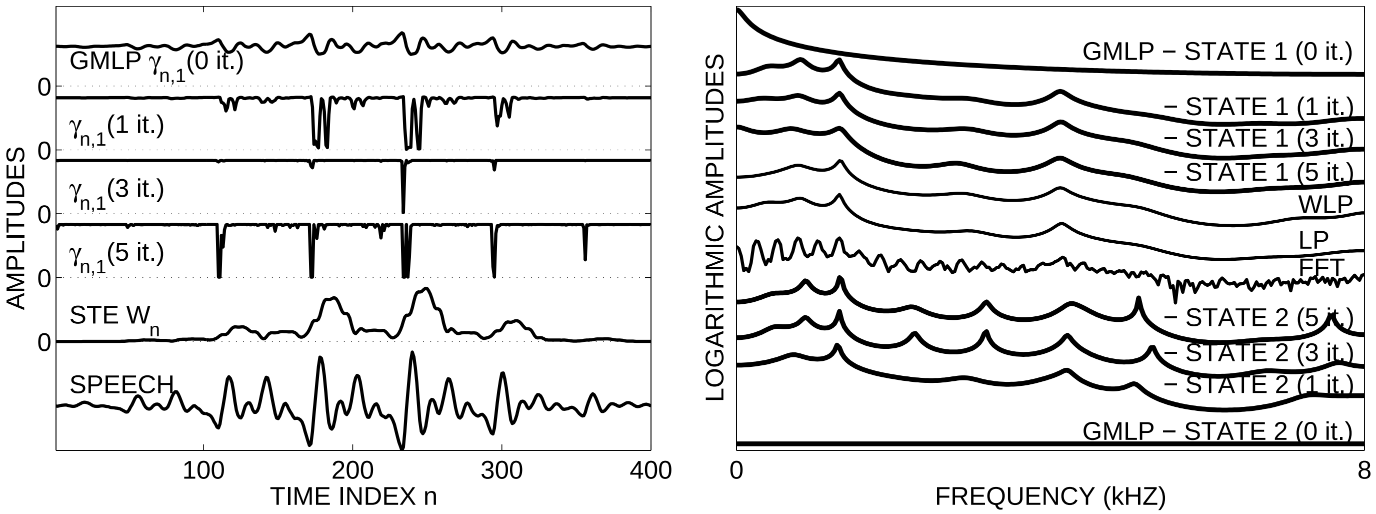

Fig. 1. Left: A Hamming-windowed speech frame (vowel from a female speaker sampled at 16 kHz) with different weighting functions. STE is the weighting scheme generally used with WLP. GMLP weights, which tend to avoid GCIs, result from iterative EM re-estimation, with the initial autoregression templates chosen according to Section II-B4. Right: Corresponding spectra of FFT, LP, WLP, and GMLP ($p = 20$), including the initial spectra of the GMLP states.

TABLE I

Equal error rates (%) for two speaker verification systems in variable test conditions, using different spectrum analyses in MFCC computation. GMLP was fixed at 3 and 5 iterations for male and female speakers, respectively, prior to the i-vector experiments, for which the methods with statistically significant improvement [31] over the other three are shown in boldface.

Training	Test	Speaker	GMM-SVM system						i-vector system			
vocal effort	vocal effort	subset	FFT	LP	WLP	GMLP (3 it.)	GMLP (5 it.)	GMLP (7 it.)	FFT	LP	WLP	GMLP
		Male	6.23	5.94	6.13	5.66	6.55	6.87	1.98	1.28	3.06	1.69
Normal	Normal	Female	8.13	8.18	8.16	6.76	7.55	8.70	3.38	3.01	3.38	3.09
	(DET5)	All	7.34	7.06	7.06	6.63	6.89	7.90	3.24	3.24	4.29	2.96
		Male	3.61	4.22	5.04	2.84	3.52	4.20	**0.21**	0.84	0.76	0.49
Normal	Low	Female	6.70	7.62	6.47	6.34	5.82	6.70	1.66	1.67	1.94	**1.11**
	(DET8)	All	5.03	6.28	5.70	5.36	5.19	5.51	1.80	1.76	1.67	1.68
		Male	7.86	8.96	8.42	7.64	8.56	7.96	3.37	3.37	4.02	3.78
Normal	High	Female	12.58	13.56	14.75	13.66	10.92	12.01	5.50	4.91	5.34	**3.82**
	(DET6)	All	10.86	11.91	12.01	10.48	9.72	10.38	5.81	5.14	6.64	4.98

TABLE II

100 × MinDCF results for the two systems evaluated analogously to Table I. Statistical testing at the MinDCF threshold (i-vector system) is done for differences in $0.5p_{\mathrm{miss}} + 0.5p_{\mathrm{fa}}$, according to which any single superior method is shown in boldface.

Training	Test	Speaker	GMM-SVM system						i-vector system			
vocal effort	vocal effort	subset	FFT	LP	WLP	GMLP (3 it.)	GMLP (5 it.)	GMLP (7 it.)	FFT	LP	WLP	GMLP
		Male	2.56	2.86	2.90	2.88	3.14	3.07	0.90	1.06	1.16	0.81
Normal	Normal	Female	3.81	3.53	3.68	3.41	3.65	3.96	1.66	1.59	1.72	1.50
	(DET5)	All	3.25	3.26	3.33	3.18	3.40	3.57	1.80	1.67	2.01	1.80
		Male	2.00	2.43	2.18	1.93	1.97	2.31	**0.22**	0.43	0.41	0.33
Normal	Low	Female	2.62	3.19	2.86	3.17	3.02	3.00	0.84	0.72	0.80	**0.58**
	(DET8)	All	2.40	2.93	2.59	2.81	2.69	2.75	1.05	0.84	1.11	0.73
		Male	4.58	4.73	4.96	4.77	4.48	4.51	1.89	1.53	1.96	1.41
Normal	High	Female	6.34	6.16	5.88	5.48	5.43	5.16	3.29	2.54	2.77	**1.73**
	(DET6)	All	5.69	5.39	5.52	5.22	4.97	4.85	3.49	2.89	3.22	2.35

IV. Conclusion

Mixture linear prediction was proposed as a stochastic version of weighted linear prediction for spectral modeling. It is given target and non-target characteristics in parameter initialization of a mixture autoregressive model prior to iterative re-estimation, which generates temporal weighting for the squared residual. In this study, the initialization was designed to focus on downweighting the effect of voiced speech excitation. In speaker verification with vocal effort (and F0) mismatch, this method significantly improved performance upon standard methods. Both the general principle and its present variant thus hold potential for further study and applications in robust signal analysis. Software implementations can be found at `http://www.acoustics.hut.fi/research/robustness/`.

REFERENCES

[1] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1295–1313, August 2013.

[2] J.-S. Liénard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *J. Acoust. Soc. Am.*, vol. 106, no. 1, pp. 411–422, July 1999.

[3] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *J. Acoust. Soc. Am.*, vol. 107, no. 6, pp. 3438–3451, June 2000.

[4] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Proc. Interspeech*, Brisbane, Australia, September 22–26 2008, pp. 609–612.

[5] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, July 2012.

[6] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 36, no. 4, pp. 433–439, April 1988.

[7] X. Fan and J. H. L. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," in *Proc. ICASSP*, Taipei, Taiwan, April 19–24 2009, pp. 4553–4556.

[8] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 411–423, February 1991.

[9] C. Ma, Y. Kamp, and L. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 1, pp. 69–81, March 1993.

[10] J. Pohjalainen, H. Kallasjoki, K. Palomäki, M. Kurimo, and P. Alku, "Weighted linear prediction for speech analysis in noisy conditions," in *Proc. Interspeech*, Brighton, UK, September 6–10 2009, pp. 1315–1318.

[11] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Process. Lett.*, vol. 17, pp. 599–602, June 2010.

[12] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[13] C. S. Wong and W. K. Li, "On a mixture autoregressive model," *Journal of the Royal Statistical Society. Series B*, vol. 62, no. 1, pp. 95–115, 2000.

[14] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, no. 2, pp. 357–384, March 1989.

[15] ——, "Analysis of time series subject to changes in regime," *Journal of Econometrics*, vol. 45, no. 1-2, pp. 39–70, July–August 1990.

[16] ——, *Time Series Analysis*. Princeton University Press, 1994.

[17] C.-J. Kim, "Dynamic linear models with Markov-switching," *Journal of Econometrics*, vol. 60, no. 1–2, pp. 1–22, January–February 1994.

[18] P. Kenny, M. Lennig, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 38, no. 2, pp. 220–225, February 1990.

[19] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. ICASSP*, Honolulu, Hawaii, April 15–20 2007, pp. 957–960.

[20] A. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. ICASSP*, Paris, France, May 3–5 1982, pp. 1291–1294.

[21] B.-H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 33, no. 6, pp. 1404–1413, December 1985.

[22] Y. Ephraim and W. J. J. Roberts, "Revisiting autoregressive hidden Markov modeling of speech signals," *IEEE Signal Process. Lett.*, vol. 12, no. 2, pp. 166–169, February 2005.

[23] B. Mesot and D. Barber, "Switching linear dynamical systems for noise robust speech recognition," *IEEE Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1850–1858, August 2007.

[24] A. P. Dempster, N. M. Laird, and D. B.Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[25] J. Pohjalainen and P. Alku, "Gaussian mixture linear prediction," in *Proc. ICASSP*, Florence, Italy, May 4–9 2014.

[26] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," University of California at Berkeley, Tech. Rep., 1998.

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

[28] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[29] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for SVM speaker recognition," in *Proc. ODYSSEY04, The Speaker and Language Recognition Workshop*, Toledo, Spain, May 31–June 3 2004, pp. 57–62.

[30] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[31] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. ODYSSEY04, The Speaker and Language Recognition Workshop*, Toledo, Spain, May 31–June 3 2004, pp. 237–244.

[32] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, January 1993.