GAUSSIAN MIXTURE LINEAR PREDICTION

Jouni Pohjalainen and Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

ABSTRACT

This work introduces an approach to linear predictive signal analysis utilizing a Gaussian mixture autoregressive model. By initializing different autoregressive states of the model to approximately correspond to the target signal and the expected type of undesired signal components, such as background noise, the iterative parameter estimation converges towards a focused linear prediction model of the target signal. Differently initialized and trained variants of mixture linear prediction are evaluated using objective spectrum distortion measures as well as in feature extraction for speech detection in the presence of ambient noise. In these evaluations, the novel analysis methods perform better than the Fourier transform and conventional linear prediction.

Index Terms— linear prediction, spectrum analysis, speech detection

1. INTRODUCTION

Linear predictive (autoregressive) modeling is used to compactly parametrize spectra of time signals, such as digital speech and audio, as all-pole filters [1]. In particular, it is widely used across the field of speech processing for analyzing and synthesizing speech signals. Conventional linear prediction (LP) analysis fits the strongest peaks of the power spectrum with an all-pole envelope. As a simple spectrum analysis method comparable to the discrete Fourier transform/fast Fourier transform (FFT), it is not designed to be particularly resistant against signal corruptions such as additive noise. Time-weighted linear predictive methods, which try to alleviate this by emphasizing certain parts of the analysis frame using various heuristic weighting schemes, have previously shown improved accuracy and robustness in several applications [2] [3] [4] [5] [6] [7].

A new probabilistic mixture decomposition approach to linear prediction is proposed in this paper. The signal is modeled as being generated by a hidden state process, where one *target* state corresponds to an all-pole filter representing the presumed target signal while the filter(s) associated with the other state(s) model expected types of signal corruptions. It is shown that the approach corresponds to a new, probabilistic form of time-weighted linear prediction. This study sets out to investigate this new approach by studying whether a suitable initialization of the state-specific all-pole filters, followed by an iterative re-estimation procedure, can tangibly improve the accuracy and/or the robustness of short-time magnitude spectrum analysis. Towards this end, the accuracy and robustness of the proposed method are evaluated and compared against standard methods by studying average spectral distances between clean and noisy speech spectra. The newly proposed approach, mixture linear prediction, is then compared against FFT and conventional LP as the basis of feature extraction of a realistic application, speech detection in acoustic monitoring of a noisy environment.

This work was supported by the EC FP7 project Simple4All (287678).

2. MIXTURE LINEAR PREDICTION

2.1. Linear prediction and weighted linear prediction

Linear predictive methods assume that the signal s_n follows a zero-mean autoregressive (AR) process $s_n = \sum_{k=1}^{p} a_k s_{n-k} + G u_n$ where G is a gain factor and u_n is the excitation signal [1]. In the z domain, this model corresponds to an all-pole filter $H(z) = G/(1 - \sum_{k=1}^{p} a_k z^{-k})$. The signal is thus assumed to be linearly predictable from its past samples as $\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}$. In standard linear prediction (LP) analysis, the predictor coefficients a_k are solved by minimization of the prediction error energy $\sum_n (s_n - \hat{s}_n)^2$. A more general formulation is weighted linear prediction (WLP) [8], which instead minimizes a time-weighted prediction error energy $E_W = \sum_n (s_n - \hat{s}_n)^2 W_n$, where the weighting function W_n is chosen to emphasize parts of the analysis frame considered most reliable. Previously, the weighting function has typically been chosen as the short-time energy (STE) of past p signal samples, $W_n = \sum_{i=1}^{p} s_{n-i}^2$ [8] [2] [4]. Standard LP, on the other hand, would follow as a special case by making $W_n = c$, a constant for all n. In any case, the coefficients of this generalized model are solved by setting $\partial E_{\mathrm{WLP}} / \partial a_j = 0$, $1 \leq j \leq p$, leading to the normal equations

$$\sum_{k=1}^{p} a_k \sum_n W_n s_{n-k} s_{n-j} = \sum_n W_n s_n s_{n-j}, \quad 1 \leq j \leq p. \quad (1)$$

In the presence of stationary background noise, STE weighting emphasizes the parts of the analysis frame with locally high signal-to-noise ratio (SNR). WLP with STE weighting and variants based on the same principle have been applied as robust spectrum analysis methods in several studies. They have shown improved accuracy and robustness in formant estimation [7] [8] as well as in feature extraction for large-vocabulary continuous speech recognition [2] [6] and text-independent speaker verification [4] [5]. Recently, it was shown in [9] that it is possible to improve the robustness of formant estimation with respect to high fundamental frequency, a typical source of difficulty in conventional LP, by designing a WLP weighting function that attenuates the overly strong contribution of the glottal source. The present study aims to improve the robustness of linear predictive analysis by developing a general method that can be made to focus on different aspects of the signal while avoiding undesired components. The proposed approach is shown to extend the previous work on weighted linear prediction.

2.2. The mixture linear prediction model

In mixture linear prediction, the signal s_n is modeled as a mixture of J autoregressive (AR) processes with conditional density

$$f(s_n | s_{n-1}, \ldots, s_{n-p}, \lambda) = \sum_{i=1}^{J} p_{n,i} \frac{1}{\sigma_i} \phi \left(\frac{u_{n,i}}{\sigma_i} \right), \quad (2)$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

where λ is the model's parameter set and $\phi(\cdot)$ is the standard normal density function so that $u_{n,i}$ is a zero-mean Gaussian white noise process, with variance σ_i^2, acting as excitation driving the AR process associated with the ith state. The AR difference equations are

$$s_n = a_{0,i} + \sum_{k=1}^{p} a_{k,i} s_{n-k} + u_{n,i}, \qquad 1 \le i \le J, \qquad (3)$$

where the $a_{0,i}$ are intercept (constant) terms. In Eq. 2, $p_{n,i} = P(q_n = i | s_{n-1}, \ldots, s_0, \lambda)$ at time n is the prior distribution of the underlying hidden state process $q_n \in \{1, \ldots, J\}$ determining which autoregressive model generates sample s_n. In the present study, q_n is considered i.i.d. and modeled as $p_{n,i} = P_i$ (see Section 2.3). Another option would be for it to follow a first-order Markov process, leading to a linear predictive hidden Markov or *Markov-switching* [10] [11] model, with a somewhat larger computational cost.

Mixture-based AR models have been extensively studied in time series analysis and econometrics using both first-order Markov [10] [11] [12] and i.i.d. [13] assumptions for the hidden state process q_n. In speech processing, they appear not to have been previously applied to low-level signal processing such as spectrum analysis. Signal models related to the current one have, however, been used in some studies as recognition models to parametrize utterances. In these studies, the hidden Markov models proposed by Poritz [14] and Juang and Rabiner [15] differ from the present model especially by applying the AR dynamics in separate frames, whereas the Markov-switching model studied by Ephraim and Roberts [16] and a related noise-aware model [17] do, similarly to the present signal model, consider each sample and its associated hidden state q_n. Vector autoregressive Gaussian mixture and hidden Markov models have also been applied on the level of feature vectors to parametrize their temporal evolution [18] [19].

2.3. Gaussian mixture linear prediction

In order to implement the method, the Gaussian mixture model (GMM) and its associated learning algorithm [20] can be extended to accommodate Gaussian autoregressive observation distributions instead of simple Gaussian distributions. A conventional (not autoregressive) univariate GMM is specified by the set of parameters $\lambda_{\mathrm{GMM}} = (P_1, \ldots, P_J, \mu_1, \ldots, \mu_J, \sigma_1^2, \ldots, \sigma_J^2)$, where P_i, μ_i and σ_i^2, $1 \le i \le J$, are the component weights (prior probability distribution of the hidden state q_n), Gaussian mean values and Gaussian variances, respectively. In contrast, the Gaussian mixture linear prediction (GMLP) model is specified by $\lambda_{\mathrm{GMLP}} = (P_1, \ldots, P_J, a_{0,1}, a_{1,1}, \ldots, a_{p,1}, a_{0,2}, \ldots, a_{p,J}, \sigma_1^2, \ldots, \sigma_J^2)$ (cf. Eq. 3). The parameters of this model can be estimated by an implementation of the iterative expectation-maximization (EM) principle [21]. Each iteration consists of an expectation (E) step followed by a maximization (M) step. In solving the model, the excitations $u_{n,i}$ are estimated as prediction residuals $e_{n,i} = s_n - a_{0,i} - \sum_{k=1}^{p} a_{k,i} s_{n-k}$.

1. In the E step, the hidden state posterior probabilities $\gamma_n(i) = P(q_n = i | s_n, \ldots, s_{n-p}, \lambda_{\mathrm{GMLP}})$ are determined as $\gamma_n(i) = \max\left(0.01, \frac{P_i (1/\sqrt{2\pi\sigma_i^2}) \exp(-e_{n,i}^2/(2\sigma_i^2))}{\sum_j^J P_j (1/\sqrt{2\pi\sigma_j^2}) \exp(-e_{n,j}^2/(2\sigma_j^2))}\right)$, i.e., a lower limit of 0.01 is imposed in this study.
2. In the M step, the component weights are re-estimated as $P_i = \frac{\sum_n \gamma_n(i)}{\sum_n 1}$ and the noise variances as $\sigma_i^2 = \frac{\sum_n \gamma_n(i) e_{n,i}^2}{\sum_n \gamma_n(i)}$. For the AR parameters $a_{k,i}$, define $x_{n,0} = 1$ (for the intercept) and $x_{n,k} = s_{n-k}$, $k > 0$, and solve the equations $\sum_{k=0}^{p} a_{k,i} \sum_n \gamma_n(i) x_{n,k} x_{n,j} = \sum_n \gamma_n(i) s_n x_{n,j}$, $0 \le j \le p$. Notably, except for the inclusion of the intercept term, the latter equations are equivalent to standard WLP (Eq. 1) weighted by the state posterior probabilities from the E step, i.e., $W_n = \gamma_n(i)$. (On the other hand, with the intercept terms, $p = 0$ would make the re-estimation procedure equivalent to that of standard GMMs [20].)

Because the EM algorithm increases the likelihood of the model with each iteration, it will converge towards a local likelihood maximum [21] whose location on the parameter hypersurface depends on the initial parameter values. A rough distinction between desired and undesired signal qualities can thus be made in choosing the initial values of the AR parameters in order to influence their final values obtained after EM re-estimation. In applying GMLP, one of the states is designated as target and the other one(s) as non-target.

3. APPLICATION TO SPEECH SPECTRUM ANALYSIS

In the present study, Gaussian mixture linear prediction with $J = 2$ is applied to robust speech analysis. The goal is to extract as target state 1 ($i = 1$ in Eq. 3) an autoregressive model that depicts the formant structure of clean speech while the non-target state 2 ($i = 2$ in Eq. 3) captures more of undesirable, noisy signal components.

To accomplish the stated goals, the target AR model is initialized with $a_{0,1} = 0$, $a_{1,1} = 0.97$ and $a_{k,1} = 0$, $2 \le k \le p$. The initial target model thus corresponds to the first-order all-pole filter $1/(1 - 0.97z^{-1})$, the inverse of the commonly used pre-emphasis filter $1 - 0.97z^{-1}$ used to compensate for the general spectral tilt of voiced speech. It can thus be viewed as a rough approximation of the low-pass spectral shape characteristic of voiced speech.

In contrast to the "speech-like" initialization of state 1, state 2 is initialized to depict a "noise" part of the signal. In the present study, this is done in three different ways. In the first variant, referred to as GMLP-0, the AR parameters of state 2 are initialized with all zeros, i.e., $a_{k,2} = 0$, $0 \le k \le p$. This gives a filter with a flat spectrum, different from the low-pass characteristics assigned to state 1, and causes the EM iteration to make the initial distinction between the states *only* based on the spectral differences, not signal amplitudes (as both intercepts are zero). Fig. 1 (left panel) shows the evolution of target and non-target spectra of GMLP-0 for a speech frame. It can be noted that the proposed method captures spectral details which are evident in the FFT spectrum but lost by conventional LP.

In the second variant, termed GMLP-H (for "High" signal value) the lagged AR parameters are also initialized with zeros, but the intercept is initialized with the largest positive signal value, i.e., $a_{0,2} = \max(s_n)$. This causes the distinction between the states to be made not only based on the spectral differences, but also by favoring (as target model) smaller amplitudes while avoiding large amplitude peaks (caused by noise or by the voice source at glottal closure instants [9]). For GMLP-H, it was found beneficial to perform one preliminary iteration of EM where only the P_i and the σ_i^2 are updated. Finally, in order to take into account the specific type of background noise encountered, the third variant GMLP-N initializes the lagged AR parameters to correspond to the spectrum of the noise, a training sample of which is assumed to be available. This is accomplished as follows: 1) the noise training signal is preprocessed and divided into frames similarly to the analyzed signal; 2) conventional LP filters are computed for each frame, converted to cepstra according to the well-known formula [22] and averaged (as averaging in the cepstral domain is perceptually meaningful [23]); 3) the conversion formula is applied in reverse direction to convert the cepstral mean to an AR model to give the parameters $a_{k,2}$, $1 \le k \le p$. For

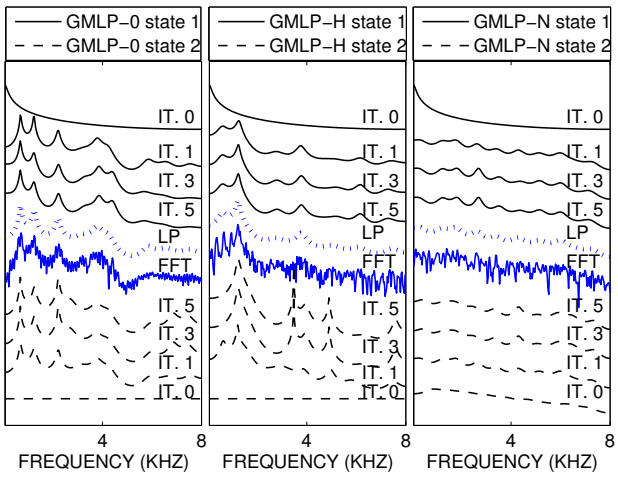

Fig. 1. Evolution of the states of mixture linear predictive models across iterations in the analysis of slightly noisy speech frames. LP and FFT spectra are shown for comparison with GMLP target model spectra after some iterations.

GMLP-N, the intercept is again initialized with zero, i.e., $a_{0,2} = 0$. Each method initializes $P_i = 0.5$ and $\sigma_i^2 = 0.01$, $1 \leq i \leq 2$.

In conventional LP, an intercept term is generally not used. If an AR process has zero mean, the intercept will also be zero [10] and can thus be omitted. For speech signals, the assumption of zero mean approximately holds using the typical analysis frame sizes of roughly 20-30 milliseconds, as speech has negligible energy at frequencies low enough to affect the mean of the frame. In GMLP, however, the inclusion of the intercept term, even if initialized with zero for each state, allows two effects. First, the target model is free to focus on any subset of the analysis frame without the constraint that those samples sum to zero. Second, as real-world noises are often of lowpass type, it is possible that the noise, which ideally would be captured by a non-target AR model, would have a non-zero mean value across the short analysis frame. However, despite the inclusion of the intercept terms in the iteration, the final AR model is chosen as $H(z) = 1/(1 - \sum_{k=1}^{p} a_{k,1} z^{-k})$, i.e., without the intercept term.

4. EXPERIMENTAL RESULTS

4.1. Evaluation against standard spectrum analysis methods using log spectral distance

Quality and robustness of spectrum models are first evaluated using the log spectral distance [23] between clean and noisy spectra as an objective quality measure. It is computed for two cases: 1) between a noisy all-pole spectrum and the corresponding clean FFT spectrum and 2) between a noisy all-pole spectrum and the corresponding clean all-pole spectrum. The speech material consists of 800 sentences from the TIMIT American English database, artificially corrupted by *factory1* and *babble* noise from the NOISEX-92 database with segmental (frame-average) signal-to-noise ratio (SNR) 20 dB, 0 dB and -20 dB. For each noise type, these distances are averaged over all non-silent 25-ms frames, excluding the sometimes silent voiced closures, according to the TIMIT phonetic transcription. The SNR-specific averaged scores are further averaged between the two noise types to produce the results shown in Fig. 2. For each of the three GMLP variants, two and eight iterations have been evaluated in order to gain an overview of the effect of this parameter.

Evidently, for some initialization methods, such as GMLP-H in this study, the number of iterations does not much change the model. For the other two methods evaluated, the number of iterations appears to function as a control parameter determining the balance between robustness (measured by degradation as compared to the same method's clean version, the vertical axis) and accuracy (measured by the distance to the clean version of FFT). On a side note, FFT itself was also evaluated in terms of robustness, but was not competitive against any of the all-pole methods, an issue which has been encountered previously in many feature extraction studies, e.g. [2] [4] [24].

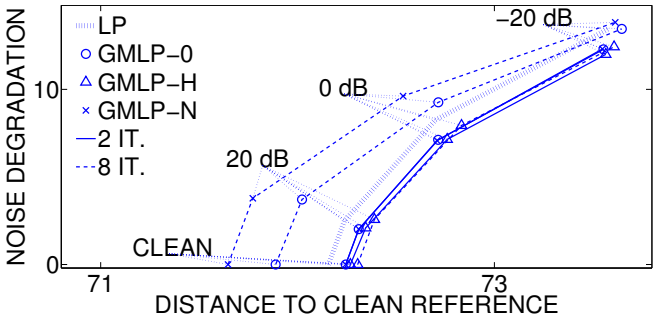

Fig. 2. Log spectral distances (in dB) from noisy all-pole spectra to clean FFT spectra (horizontal axis) and to the clean version of the same all-pole method (vertical axis) plotted as a trajectory in two dimensions with signal-to-noise ratio (SNR) as parameter.

4.2. A system for speech detection in environment monitoring

We consider speech activity detection in acoustic environment monitoring as a practical application for the method. In security-oriented applications, such systems typically aim to detect abnormal sounds which can include shouts, screams, gunshots, explosions, banging sounds and non-neutral speech [25] [26] [27] [28] and may require speech detection capability in order to detect non-neutrality or to characterize speakers by paralinguistic analysis [29]. In secluded areas where speech activity typically does not occur, speech can be a target event of interest in its own right. Besides surveillance applications, long-term speech event detection can be employed with, e.g., various context-aware user interfaces or simply for logging the long-term activity patterns in a particular environment [30].

Compared to typical voice activity detection (VAD), which is usually performed to assist speech coding [31], automatic speech recognition or speaker recognition [32] and a detection decision is generally required within frames of speech, the environment monitoring application has two special characteristics. First, the detection decisions are only required on a coarse time scale of seconds. Thus, the system has some freedom to focus on detection performance and noise robustness by using a longer analysis window and increasing the detection delay. Second, both the recording environment and the transmission channel are known, as the recording equipment is assumed to be installed in a fixed position or at least to stay in the same location for long time periods. Channel normalization is thus not an issue and it also becomes possible to train statistical models off-line to represent the known noise environment. To address these considerations, we adapt our earlier system for detecting shouted speech in noise [25] which is based on statistical classification of mel-frequency cepstral coefficient (MFCC) feature vectors.

The input audio signal, sampled at 16 kHz and pre-emphasized with $H(z) = 1 - 0.97z^{-1}$, is processed in analysis frames of 25 ms taken every 10 ms. The detection decisions are made once per second using an analysis block length of five seconds. MFCCs are obtained for each frame using a usual processing chain [33]: 1) squared magnitude spectrum computation; 2) weighted mel-filterbank band energy computation (here, we use 40 triangular filters spaced evenly on the mel scale); 3) logarithm and 4) discrete cosine transform. In step 1, similarly to earlier studies, we substitute all-pole methods as alternative to the standard FFT for magnitude spectrum computation. Twelve MFCCs, excluding the zeroth one, constitute the feature vector. Inclusion of frame energy or delta coefficients was experimented with but was not found to improve the performance.

After feature extraction, high-energy frames are selected. This is accomplished by computing the logarithmic energy of each frame within the five-second block and clustering these 500 energy values E_n into two clusters by k-means [33] after initializing the cluster means with $\min(E_n)$ and $\max(E_n)$ [25]. Only the frames ending up in the cluster initialized with $\max(E_n)$ are selected for further processing, i.e., either to serve as training data (in the training phase, overlapped frame selection decisions are averaged) or to contribute to the detection decision in the classification phase. This approach focuses on the locally most energetic frames that have the highest SNR in the case of stationary background noise.

Gaussian mixture models (GMMs) with eight components and diagonal covariance (using more components did not produce better results) are trained using ten iterations of EM for GMMs [20] to represent M different *speech* classes and *non-speech*. In this study, the latter class corresponds to background noise. Before training, the component weights of each GMM are initialized with uniform distributions, the variance parameters by 0.1 times the global variances of the features, and the mean parameters by a heuristic selection approach [34]. The detector computes averaged logarithmic likelihoods, denoted by $L_{\text{speech,1}}, \ldots, L_{\text{speech,M}}$ and $L_{\text{non-speech}}$, of the selected feature vectors having been produced by each GMM. Speech is decided to be present if $L = \max_m(L_{\text{speech,m}}) - L_{\text{non-speech}} > T$, where T is a threshold for the log likelihood ratio L.

4.3. Results for speech detection in noise

The speech material consists of 24 Finnish sentences, each spoken by 11 male and 11 female speakers [25]. The detection system is trained using clean speech but the test speech is artificially corrupted by noise. The *factory1* and *factory2* noises from NOISEX-92 are considered, the former of which is less stationary and contains obvious impulsive and transient sounds. For each test utterance, a pure noise segment of equal length is also classified. In order to ensure speaker independence, the evaluation is carried out as 22-fold cross validation where each speaker in turn is chosen as the test speaker and the material from the other 21 speakers is used for training the models. The experiment is first carried out for detecting normal speech only ($M = 1$) and then, because the complete material contains the same sentences spoken both by normal voice and by shouting, for detecting speech with variable vocal effort, i.e., by including also shouted speech in the training and test material ($M = 2$).

Table 1 shows the equal error rates (EER) with *factory1* and *factory2* noise. The statistical significance of the differences between methods was evaluated using a significance test appropriate for detection systems [35]. The "dependent-case" version of this test was employed, as all the detections use the same analysis block division and original speech material. With normal speech, each spectrum analysis method achieves perfect detection above and at 0 dB SNR (not shown) with both noise types. The GMLP methods generally perform best, and in many cases achieve statistically significant improvement over both baselines FFT and LP, as denoted by boldface. GMLP initialization obviously has an effect on performance but each evaluated GMLP method performs at least adequately in comparison to the baselines. Compared to the markedly nonstationary *factory1* noise, *factory2* is noticed to be easier and also gets more help from modeling by GMLP-N. As expected, inclusion of shouted speech generally improves the detection scores in a given noise scenario.

Table 1. *EER scores (%) for speech detection in two types of noise using FFT, LP and three GMLP variants as base spectrum analysis methods in MFCC computation. Detection scores are also shown for the case in which both the training and test material contain both normal and shouted speech. In this case, scores for the best number of iterations among 2, 5 and 8 are shown for each GMLP method. The GMLP scores that are statistically significantly better than both FFT and LP in the corresponding case are indicated in boldface.*

	factory1			*factory2*		
	SNR (dB)					
Method	-5	-10	-15	-5	-10	-15
Normal speech only						
FFT	2.5	18.4	22.8	0.0	4.7	21.0
LP	3.5	19.8	32.0	0.2	6.0	21.5
GMLP-0 (2 it.)	**1.4**	**14.4**	**21.7**	0.2	5.4	21.5
(5 it.)	**1.1**	**7.6**	26.1	0.2	6.0	22.0
(8 it.)	2.5	**11.4**	35.0	0.2	4.7	21.5
GMLP-H (2 it.)	2.2	**13.6**	22.2	0.0	4.7	21.4
(5 it.)	**0.9**	**9.5**	24.1	0.2	6.2	21.5
(8 it.)	**1.4**	**16.1**	28.8	0.3	7.0	22.0
GMLP-N (2 it.)	**1.9**	**16.0**	**21.2**	0.2	5.7	21.5
(5 it.)	**0.8**	**9.7**	26.1	0.0	**2.8**	**16.9**
(8 it.)	**1.6**	**11.9**	31.0	0.0	**2.2**	**16.3**
Normal and shouted speech						
FFT	1.4	9.6	14.9	0.0	2.6	12.2
LP	0.8	9.1	14.2	0.0	3.4	13.5
GMLP-0 (2 it.)	**0.5**	**7.6**	**12.9**	0.0	3.4	13.2
GMLP-H (2 it.)	0.8	9.2	**11.9**	0.0	2.6	12.3
GMLP-N (5 it.)	0.8	**6.5**	13.4	0.0	**1.7**	**9.3**

5. CONCLUSIONS

Mixture linear prediction was described and applied to speech spectrum analysis with a focus on noise robustness. Autoregressive parameters associated with two states of a mixture model were initialized to broadly characterize, respectively, the desired target signal and undesired, noisy components. Three such variants of the proposed method were evaluated in additive-noise conditions, both using an objective quality measure, the log-spectral distance, and in a practical application, feature extraction for speech detection in acoustic environment monitoring. The described method was able to outperform conventional methods in both experiments. Within the scope of the evaluations, it is therefore noted that the target state generally converges towards a useful all-pole model during the course of EM re-estimation of the model parameters. While the results were similar across different initialization methods, AR parameter initialization and the number of training iterations were also observed to have a potentially large effect on the resulting all-pole model. Initialization approaches are thus among the topics of future study, as well as the structure of the mixture model, modifications to training and the use of mixture-based linear prediction in new applications.

6. REFERENCES

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

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

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

[4] 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, no. 6, pp. 599–602, 2010.

[5] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proc. Interspeech*, Makuhari, Japan, September 2010.

[6] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo, "Noise robust feature extraction based on extended weighted linear prediction in LVCSR," in *Proc. Interspeech*, Florence, Italy, August 2011.

[7] D. Gowda, J. Pohjalainen, M. Kurimo, and P. Alku, "Robust formant detection using group delay function and stabilized weighted linear prediction," in *Proc. Interspeech*, Lyon, France, August 2013.

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

[9] 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, 2013.

[10] J. D. Hamilton, *Time Series Analysis*, Princeton University Press, 1994.

[11] C.-J. Kim, "Dynamic linear models with Markov-switching," *Journal of Econometrics*, vol. 60, pp. 1–22, 1994.

[12] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, pp. 357–384, 1989.

[13] C. S. Wong and W. K. Li, "On a mixture autoregressive model," *Journal of the Royal Statistical Society. Series B*, vol. 62, pp. 95–115, 2000.

[14] A.B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. ICASSP*, Paris, France, May 1982.

[15] 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, 1985.

[16] 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, 2005.

[17] 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, 2007.

[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, 1990.

[19] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. ICASSP*, Honolulu, Hawaii, April 2007.

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

[21] 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, pp. 1–38, 1977.

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

[23] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 24, no. 5, pp. 380–391, 1976.

[24] F. de Wet, B. Cranen, J. de Veth, and L. Boves, "A comparison of LPC and FFT-based acoustic features for noise robust ASR," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[25] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: human and machine," *J. Acoust. Soc. Am.*, vol. 133, no. 4, pp. 2377–2389, 2013.

[26] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Int. Conf. AVSS*, London, UK, September 2007.

[27] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP J. on Audio, Speech, and Music Processing*, 2009.

[28] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in *Proc. IEEE WASPAA*, New Paltz, USA, October 2005.

[29] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - state-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, pp. 4–39, 2013.

[30] A. Härmä, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME 2005)*, Amsterdam, The Netherlands, July 2005.

[31] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[32] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[33] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

[34] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized Lloyd iteration," *IEEE Signal Process. Lett.*, vol. 1, pp. 144–146, 1994.

[35] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. ODYSSEY04*, Toledo, Spain, June 2004.