ROBUST SPEECH PROCESSING USING ARMA SPECTROGRAM MODELS

Sriram Ganapathy

IBM T.J Watson Research Center, Yorktown Heights, NY, USA.

ABSTRACT

Speech applications in noisy and degraded channel conditions continue to be a challenging problem especially when there is a mismatch between the training and test conditions. In this paper, a robust speech feature extraction scheme is developed based on autoregressive moving average (ARMA) modeling that emphasizes high energy regions of the signal with a data driven modulation filter. The peak preserving ability of two dimensional autoregressive (AR) models is used to emphasize the high energy regions in the spectro-temporal domain. The modulation filtering property is achieved by moving average (MA) modeling. The ARMA spectrograms are used to derive features for speech recognition in the Aurora-4 database. In these experiments, the ARMA model features provide significant improvements (relative improvements of 15%) compared to other robust features. Furthermore, the robustness of these features is also verified for language identification (LID) of highly degraded radio channel speech. Here, the ARMA approach achieves relative improvements of up to 20% over the baseline features.

***Index Terms*—** Robust Feature Extraction, ARMA Modeling, Speech Recognition, Language Identification.

1. INTRODUCTION

Even with several advancements in the practical application of speech technology, the performance of the state-of-the-art systems remain fragile in high levels of noise and other environmental distortions. On the other hand, various studies on the human auditory system have shown good resilience of the system to high levels of noise and degradations [1]. This information shielding property of the auditory system may be largely attributed to the signal peak preserving functions performed by the cochlea and the spectro-temporal modulation filtering performed in the cortical stages. In this paper, we attempt to emulate some of these properties for robust feature extraction.

One common solution to overcome the performance degradation in noisy conditions is the use of multi-condition training [2] where the acoustic models are trained using data from the target domain. However, in a realistic scenario it is not always possible to obtain reasonable amounts of training data from all types of noisy environments. Therefore, there is a need to attain noise robustness either at the front-end signal analysis or at the statistical modeling stage. The goal of this paper is to address the robustness issues in feature extraction.

Various techniques like spectral subtraction [3], Wiener filtering [4], power bias subtraction [5] and missing data reconstruction [6] have been proposed for noisy speech recognition scenarios. Feature compensation techniques have also been used in the past like feature warping [7], RASTA processing [8] and cepstral mean subtraction (CMS) [9]. ARMA filtering of cepstral features have also been proposed for speech recognition [10]. In many of these techniques, there is an assumption of additive or convolutive noise model. However, in a realistic scenario, it is not always possible to characterize the noise model especially for non-linear channel distortions like radio channels [11]. In this paper, we propose to develop a robust front-end which is devoid of any noise model.

In general, an autoregressive (AR) modeling approach represents high energy regions with good modeling accuracy [12, 13]. One dimensional AR modeling of signal spectra is widely used for feature extraction of speech in the form of perceptual linear prediction (PLP) [14]. The one dimensional temporal AR model has been proposed in the past using frequency domain linear prediction [15, 16]. Recently, it was shown that 2-D AR modeling with modulation filtering can generate robust speech representations [17]. In this paper, we extend this approach using autoregressive moving average (ARMA) spectrogram modeling. The ARMA process is a generalization of AR modeling and can estimate band-pass characteristics while the AR modeling typically estimates low-pass characteristics [18]. In our case, the ARMA modeling is applied on the sub-band discrete cosine transform (DCT) components for estimating temporal envelopes. The ARMA filtered envelopes are used to obtain a spectrographic representation by short-term integration. Then, linear prediction based spectral smoothing is applied on this spectrogram and used for speech/language recognition in noisy conditions.

The automatic speech recognition (ASR) experiments are performed on the noisy speech from the Aurora-4 database using a deep neural network (DNN) acoustic model [6]. The results from these experiments indicate that the ARMA modeling approach provides significant improvements (relative improvements of 15%) over other noise robust front-ends. Furthermore, language identification (LID) experiments performed on highly degraded radio channel speech [11] confirm the generality of the proposed features for a wide range of noise conditions.

The rest of the paper is organized as follows. In Sec. 2, we outline the proposed ARMA spectrogram derivation for feature extraction. Sec. 3 describes the ASR experiments using the proposed front-end. In Sec. 4, we describe our experimental setup and the results for a language recognition task. In Sec. 5, we conclude with a brief discussion of the proposed front-end.

2. ARMA SPECTROGRAM ESTIMATION

2.1. Background

In this subsection, we briefly highlight the difference in the problem formulation of AR and ARMA models.

This work was supported in part by Contract No. D11PC20192 DOI/NBC under the RATS program. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

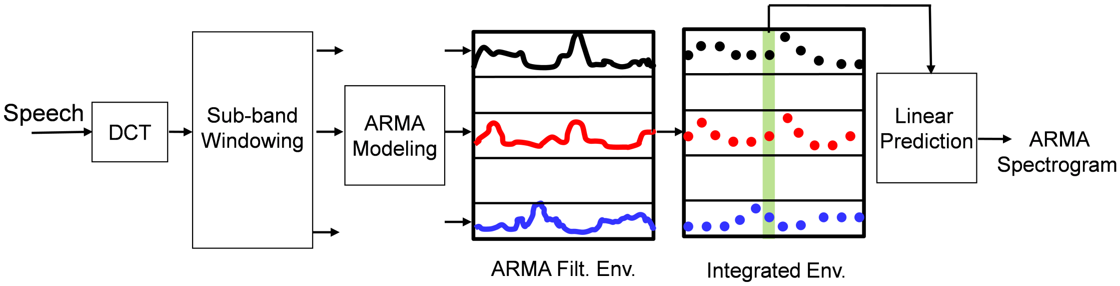

Fig. 1. Spectrogram estimation using temporal ARMA modeling.

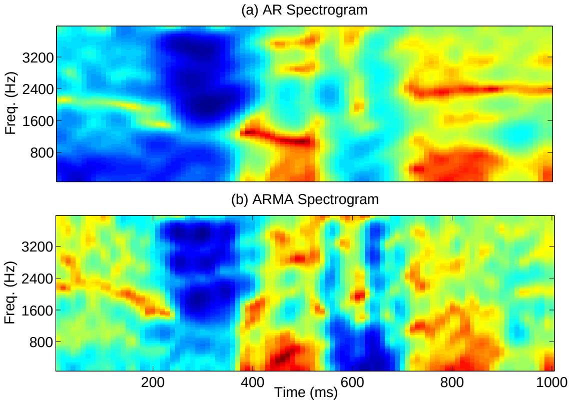

Fig. 2. Comparison of AR ($p = 40$) with ARMA ($p = 40, q = 6$) spectrogram for a 1000ms portion of speech signal.

2.1.1. *AR Modeling*

Autoregressive (AR) modeling of short-term spectrum is widely used in speech and audio signal processing for about four decades now [12, 13]. Let $x[n]$ denote the input signal for $n = 0, \ldots, N-1$. The time domain LP model is formulated to identify the set of coefficients $a_j, j = 1, \ldots, p$ such that $\sum_{j=1}^{p} a_j x[n-j]$ approximates $x[n]$ in a least square sense [12], where p denotes the model order. Let $\mathbf{r}_x[\tau]$ denote the autocorrelation sequence for time domain signal $x[n]$ with lag τ ranging from $-N+1, \ldots, N-1$.

$$r_x[\tau] = \frac{1}{N} \sum_{n=|\tau|}^{N-1} x[n]x[n-|\tau|] \tag{1}$$

Let $\hat{x}[n]$ denote the zero-padded signal $\hat{x}[n] = x[n], \quad n = 0, .., N-1$ and $\hat{x}[n] = 0, \; for \; n = N, .., 2N-1$. The relation between the power spectrum of the signal $P_x[k] = |\hat{X}[k]|^2$ and the autocorrelation $\mathbf{r}_x[\tau]$ is given by the Fourier relation,

$$P_x[k] = \mathcal{F}\big[r_x[\tau]\big] \tag{2}$$

where $\hat{X}[k]$ is the discrete Fourier transform (DFT) of the signal $\hat{x}[n]$ for $k = 0, \ldots, 2N-1$. This relation is used in the AR modeling of the power spectrum of the signal [13]. The time domain linear prediction (TDLP) refers to the use of time domain autocorrelation sequence to solve the linear prediction problem. The optimal set of a_j along with the variance of prediction error G with $a_0 = 1$ provides an AR model of the power spectrum,

$$\hat{P}_x[k] = \frac{G}{|\sum_{j=0}^{j=p} a_j e^{-i2\pi jk}|^2} \tag{3}$$

The frequency domain linear prediction (FDLP) model was proposed by Kumaresan [15]. This was reformulated by Athineos and Ellis [16] using matrix notations and the connection with DCT sequence is established. A simplified derivation without using matrix notations is provided in [19]. In the FDLP model, the problem is formulated to identify the set of coefficients $a_j, j = 1, \ldots, p$ such that,

$$X[k] = \sum_{l=1}^{p} a_l X[k-l] + U[k], \tag{4}$$

where $X[k]$ are DCT components of the signal $x[n]$ and p denotes the FDLP model order.

The fundamental relationship underlying the FDLP model is that the auto-correlation of the DCT signal and the squared magnitude of the analytic signal (Hilbert envelope) are Fourier transform pairs. This is exactly analogues to the relation in Eq. 2. In other words, AR modeling of Hilbert envelope can be achieved by linear prediction of DCT components.

2.1.2. *ARMA Modeling*

In the proposed framework, we use the DCT components in an ARMA modeling framework to estimate the sub-band envelope. The ARMA model applied on DCT components $X[k]$ is the identification of the set of coefficients $a_l, l = 1, \ldots, p$ and $b_m, m = 1, \ldots, q$ such that,

$$X[k] = \sum_{l=1}^{p} a_l X[k-l] + \sum_{m=0}^{q} b_m U[k-m], \tag{5}$$

where p, q denote the model order of the AR and MA components and $U[k]$ denotes zero mean white noise signal. The AR model is a specific case of the ARMA model with $b_m = 0, \; for \; m > 0$. The ARMA envelope is given by,

$$\hat{E}_x[n] = \frac{|\sum_{m=0}^{q} b_m e^{-i2\pi mn}|^2}{|\sum_{l=0}^{p} a_l e^{-i2\pi ln}|^2} \tag{6}$$

Comparing Eq. 3 and Eq. 6, we find that ARMA envelope is the AR envelope multiplied by a finite impulse response (FIR) filter provided by the MA modeling. Since the estimation is applied to obtain the sub-band Hilbert envelope, the MA filter acts as a modulation filter over long temporal regions of signal. Thus, ARMA modeling combines AR estimation with a data-driven modulation filter. In our estimation, we use gain normalized ARMA envelopes ($a_0 = 1$ and $b_0 = 1$).

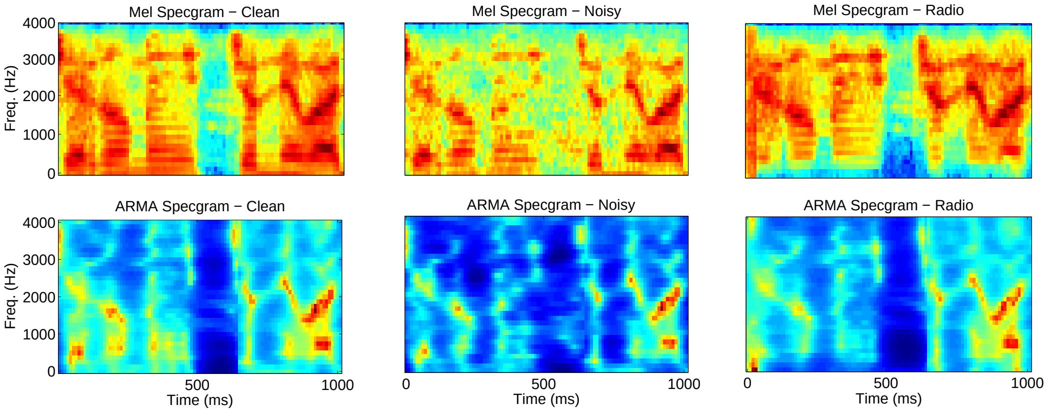

Fig. 3. Comparison of mel spectrogram and ARMA spectrogram for clean speech, noisy speech (babble noise at 10 dB SNR) and radio channel speech (channel C).

2.2. Feature Extraction

The block schematic of the proposed approach for feature extraction is shown in Fig. 1. Long segments of the input speech signal (1000ms of non-overlapping windows) are transformed using DCT. The full-band DCT signal is windowed into a set of overlapping sub-bands. The ARMA modeling is applied on the sub-band DCT components to estimate the sub-band envelope Eq.(6). In our ARMA estimation, the numerator and denominator are estimated separately to reduce the computational complexity.

The sub-band ARMA envelopes are integrated with a Hamming window over a 25 ms window with a 10 ms shift. The integration in time of the sub-band envelope yields an estimate of the short-term power spectrum. For each 25 ms frame, these power spectral estimates are transformed to temporal autocorrelation estimates using inverse Fourier transform and used for time domain linear prediction (TDLP). This gives the spectrally smoothed ARMA spectrogram.

Although ARMA modeling can also be applied to spectral domain, we apply ARMA model only in the temporal domain for this work. In Fig.2, we compare the spectrographic representation from AR and ARMA modeling. As seen here, the AR modeling results in a smooth representation which emphasizes only the high energy regions of the signal. The ARMA modeling on the other hand, enhances the changes in the signal energy while suppressing the constant regions (band-pass modulation filtering) in addition to modeling the signal peaks. The tradeoff between these two modeling properties can be controlled in the ARMA model by varying the model order values $[p, q]$. Unlike the previous techniques like RASTA [8], the modulation filtering in ARMA modeling is data driven. This modulation filtering property of the ARMA model provides good noise robustness properties as shown in the experiments.

In Fig. 3, we compare the spectrographic representation of the speech signal in three conditions - clean speech, noisy speech (additive babble noise at 10 dB signal-to-noise ratio (SNR)) and radio channel speech (from channel C in the RATS database [11]). The plots compare the representation from the conventional mel frequency analysis with the ARMA representation. As seen here, the proposed approach yields a representation focussing on important regions of the clean signal. For the degraded conditions, the representation provides a good match with the clean signal suppressing the effects of noise.

Table 1. Word error rate (%) in Aurora-4 database with clean training for various feature extraction schemes.

Cond.	MFBE	ETSI	PNFBE	AR	ARMA
Clean Same Mic					
Clean	3.1	3.1	**2.8**	3.1	3.0
Clean Diff. Mic					
Clean	14.9	14.8	**11.3**	**11.3**	11.7
Additive Noise Same Mic					
Airport	23.6	13.6	17.6	15.4	**13.7**
Babble	20.7	14.1	15.9	15.2	**13.0**
Car	8.0	8.7	5.9	5.6	**5.0**
Restaurant	26.3	19.4	21.9	19.1	**17.3**
Street	19.8	18.3	16.9	14.8	**13.6**
Train	20.8	16.9	16.0	14.9	**14.5**
Avg.	19.9	15.2	15.7	14.2	**12.9**
Additive Noise Diff. Mic					
Airport	41.5	29.9	35.6	31.2	**29.5**
Babble	38.4	31.3	34.3	31.1	**29.6**
Car	25.8	23.9	20.7	**17.8**	18.4
Restaurant	41.3	34.0	37.4	32.4	**31.1**
Street	38.1	33.5	33.1	29.2	**28.3**
Train	37.3	32.1	31.7	29.2	**29.1**
Avg.	37.1	30.8	32.1	28.5	**27.7**

3. SPEECH RECOGNITION EXPERIMENTS

We perform a set of automatic speech recognition experiments in the Aurora4 database [20] using a deep neural network (DNN) hybrid system [6]. We use the clean training setup which contains 7308 clean recordings (14h) for training the acoustic models. The system uses a tri-gram language model with 5k vocabulary size. The test data consist of 330 recordings each from 14 conditions which include clean testing with same microphone, clean testing with different microphone, 6 additive noise conditions which include airport, babble, car, restaurant, street and train noise at $5-15$ dB signal-to-noise ratio (SNR) and 6 conditions with the combination of additive and channel noise.

We experiment with various feature extraction methods for the DNN-ASR system namely - mel filter bank energies (MFBE), power

Table 2. LID performance (in terms of EER (%)) for various feature techniques for 120s, 30s and 10s duration.

	120s				**30s**				**10s**			
Cond.	MFCC	MVA	PNCC	ARMA	MFCC	MVA	PNCC	ARMA	MFCC	MVA	PNCC	ARMA
Chn. A	21.0	12.5	15.0	**8.1**	21.0	13.3	17.5	**11.0**	24.5	20.0	23.6	**16.6**
Chn. C	14.5	16.6	13.9	**13.2**	13.8	15.4	10.9	**10.5**	20.0	22.1	19.4	**17.7**
Chn. D	18.5	16.6	13.1	**12.2**	22.0	19.1	**16.1**	17.3	24.3	22.9	**19.5**	20.2
Chn. F	12.4	19.9	7.7	**5.5**	11.5	16.7	10.1	**6.9**	17.3	23.2	14.5	**12.5**
Avg.	16.6	16.4	12.4	**9.7**	17.1	16.1	13.7	**11.4**	21.3	22.1	19.3	**16.7**

Table 3. Description of RATS radio communication channels [11].

Channel	Characteristic
A	Receiver 50kHz offset
C	Receiver 3kHz offset
D	Frequency Shift
F	Spread Spectrum

normalized filter bank energies (PNFBE) [5] and ETSI [4]. All these features use a 21 frame context with utterance based mean variance normalization. We also compare the ARMA filtering method with the AR modeling technique [21]. For these features, we use 14 modulation components from each mel-band obtained by a DCT on 200 ms windows of sub-band envelopes. The modulation components are spliced with their frequency derivatives to form the input features for the DNN. For the ARMA spectrogram estimation, we use $p = 40, q = 6$ poles per second per sub-band. We also use a compression factor of 0.2 on the MA part for envelope computation.

For all the acoustic features, the ASR model consists of a DNN with 4 hidden layers of 1024 activations and uses context dependent phoneme targets obtained from an initial alignment using a hidden-Markov-model-GMM system. The DNNs are generatively pre-trained with a restricted Boltzmann machine (RBM) trained on the acoustic features. The DNN training and ASR setup are obtained from the Kaldi toolkit [22]. The performance of the ASR system is measured in terms of word error rate (WER).

The ASR results for various feature processing schemes is shown in Table 1. Among the baseline features, the ETSI features provide the best ASR performance on the noisy conditions. The AR modeling approach improves the performance on all the noisy conditions compared to the ETSI features. The ARMA modeling combines the benefits of AR modeling with the modulation filtering provided by MA modeling. The ARMA model based features provide the best performance in the noisy conditions of the Aurora-4 task. On the noisy conditions with the same microphone, the ARMA model achieves an average performance improvement of 10% relative compared to AR model and 15% compared to the ETSI features. For the noisy conditions with different microphone, the ARMA model provides an average relative improvement of 10% over the ETSI features. Furthermore, the improvement obtained by ARMA model features over the AR model features shows the benefits of combining the AR model with the MA modulation filter.

4. LANGUAGE RECOGNITION EXPERIMENTS

The development and test data for the LID experiments use the LDC releases of phase-I RATS LID evaluation [11]. This consists of speech recordings from previous NIST-LRE clean recordings as well as other RATS clean recordings passed through eight noisy radio communication channels. Each channel induces a degradation mode to the audio signal based on its device non-linearities, carrier modulation types, network parameter settings etc [11].

The five target languages are Arabic, Farsi, Dari, Pashto and Urdu. In order to investigate the effects of an unseen communication channel (not seen in training), we divide the eight channels to two groups - channels B,E,G,H used in the training and the channels A,C,D,F used in testing. This division of channels is done to target the realistic application of these systems where the noise and channel characteristics of the test data are not available during training. The description of the four test channels is given in Table 3.

The training data consist of $24,123$ recordings with 270 hours of data from each of the four noisy communication channels (B,E,G,H) and the test set consists of $7,164$ recordings with about 15 hours of data from each of the four target channels of interest (A,C,D,F). The training and test recordings consist of 120s, 30s and 10s speech segments. The speech features are processed with feature warping [7] and are used to train a Gaussian mixture model-Universal background model (GMM-UBM) with 1024 mixture components. Then, an i-vector projection model of 300 dimensions is trained [23]. The back-end classifier is a multi-layer perceptron (MLP) trained with the i-vectors as the input and the corresponding language labels as the targets [24]. The MLP has 2000 hidden units and is trained with a cross-entropy cost function. The performance of the LID system is measured in terms of equal error rate (EER).

We experiment with various feature extraction schemes in the LID system like - MFCC features, MVA features [10], PNCC features [5] and the proposed ARMA modeling approach. The results for the LID experiments for various features is shown in Table. 2. As seen here, the PNCC features provide the best baseline performance on these highly degraded noisy channels. The proposed ARMA features provide significant improvements over the other features considered here for most of the channel conditions (except channel-D in 30s and 10s duration). We obtain an average relative performance improvement of $15 - 20\%$ for the ARMA filtering approach over the PNCC baseline. These results are in conjunction with the ASR results and indicate the consistency of the proposed approach for variety of speech applications involving various types of artifacts like additive and convolutive noise as well as non-linear radio channel distortions.

5. SUMMARY

In this paper, we have proposed an ARMA model of spectrogram for noise robust feature extraction. The ARMA model is applied on sub-band DCT components to estimate the temporal envelopes and it combines the peak estimation properties of AR approach along with the modulation filtering property of MA modeling. We perform several speech recognition and language identification experiments in noisy and degraded channel conditions. In these experiments, the proposed features provide significant improvements compared to various other noise robust front-ends and exhibit good generalization to a wide variety of acoustic distortions.

6. REFERENCES

[1] S. Greenberg, W. Ainsworth, A. Popper, and R. Fay, *Speech processing in the auditory system*, vol. 18, Springer, 2004.

[2] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] "ETSI ES 202 050 v1.1.1 STQ : Distributed speech recognition : Advanced Front-End : Compression algorithms," 2002.

[5] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *ICASSP*, 2012.

[6] M. L Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *ICASSP*. IEEE, 2013, pp. 7398–7402.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Speaker Odyssey, Speaker Recognition Workshop*, 2001.

[8] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[9] A.E. Rosenberg, C.H. Lee, and F.K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Third International Conference on Spoken Language Processing*, 1994.

[10] C. P. Chen and Jeff A Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[11] K. Walker and S. Strassel, "The RATS Radio traffic collection system," in *Odyssey Speaker and Language Recognition Workshop*, 2012.

[12] B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of Acoustical Society of America*, vol. 47, pp. 637–655, 1970.

[13] J. Makhoul, "Linear prediction: A tutorial review," *Proc. of the IEEE*, vol. 63, pp. 561–580, 1975.

[14] H. Hermansky, "Perceptual linear predictive analysis of speech," *Journal of Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[15] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *Journal of Acoustical Society of America*, vol. 105, pp. 1912–1920, 1999.

[16] M. Athineos and D.P.W. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Tran. on Sig. Proc.*, vol. 55, pp. 5237–5245, 2007.

[17] S. Ganapathy and M. Omar, "Auditory motivated front-end for noisy speech using spectro-temporal modulation filtering," *Journal of Acoustical Society of America - Express Letters*, vol. 136, pp. 343–349, 2014.

[18] D. B. Percival, *Spectral analysis for physical applications*, Cambridge University Press, 1993.

[19] S. Ganapathy, "Signal analysis using autoregressive models of amplitude modulation," *PhD Thesis, Johns Hopkins University*, 2012.

[20] Hans-Gunter Hirsch and David Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop*, 2000.

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

[22] D. Povey et al., "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011, pp. 1–4.

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

[24] P. Matejka et al., "Patrol team language identification system for DARPA RATS P1 evaluation.," in *INTERSPEECH*, 2012.