Time-Varying Autoregressions for Speaker Verification in Reverberant Conditions

Ville Vestman[1]*, Dhananjaya Gowda*[2]*, Md Sahidullah*[1]*, Paavo Alku*[3]*, Tomi Kinnunen*[1]

[1]School of Computing, University of Eastern Finland, Finland
[2]DMC R&D Center, Samsung Electronics, Seoul, Korea
[3]Department of Signal Processing and Acoustics, Aalto University, Finland

{vvestman, sahid, tkinnu}@cs.uef.fi, d.gowda@samsung.com, paavo.alku@aalto.fi

Abstract

In poor room acoustics conditions, speech signals received by a microphone might become corrupted by the signals' delayed versions that are reflected from the room surfaces (e.g. wall, floor). This phenomenon, reverberation, drops the accuracy of automatic speaker verification systems by causing mismatch between the training and testing. Since reverberation causes temporal smearing to the signal, one way to tackle its effects is to study robust feature extraction, particularly based on long-time temporal feature extraction. This approach has been adopted previously in the form of 2-dimensional autoregressive (2DAR) feature extraction scheme by using frequency domain linear prediction (FDLP). In 2DAR, FDLP processing is followed by time domain linear prediction (TDLP). In the current study, we propose modifying the latter part of the 2DAR feature extraction scheme by replacing TDLP with time-varying linear prediction (TVLP) to add an extra layer of temporal processing. Our speaker verification experiments using the proposed features with the text-dependent RedDots corpus show small but consistent improvements in clean and reverberant conditions (up to 6.5%) over the 2DAR features and large improvements over the MFCC features in reverberant conditions (up to 46.5%).

Index Terms: speaker recognition, autoregressive modeling, autocorrelation domain time-varying linear prediction

1. Introduction

An automatic speaker verification system is said to be *robust* if it tolerates external distortion caused by environmental noise or transmission channel, or internal signal variation caused by, for example, different speaking styles. A large part of previous speaker verification studies have focused on robustness with respect to noise (e.g. [1]), while robustness with respect to varying room acoustics conditions has remained less studied. Room acoustics, however, might have a large effect on the accuracy of a speaker recognition system: If the training and testing data are recorded in different room environments, there will be a mismatch between the training and testing which will degrade recognition accuracy. Room acoustics affect speech signals particularly in the form of *reverberation* [2]: the signal received by a microphone becomes a sum of the direct component and its delayed versions that arrive at the microphone after being reflected from the surfaces (e.g. walls, floor, ceiling) of the room. Mismatch caused by reverberation can in principle be tackled by modifying the speaker verification front-end or back-end. In our view, advances on both sides are necessary to reach the best possible performance. On the back-end side, techniques such as multi-condition training [3], where multiple speaker models for different reverberation conditions are created for each speaker, can be utilized. In the current study, however, we focus on the system front-end by studying features that aim at reducing the mismatch caused by reverberation.

Speech features robust to reverberation have been previously investigated in a few speaker verification studies. In [4], the use of *locally normalized cepstral coefficients* (LNCCs) was studied. LNCC features modify the conventional MFCC features by using an additional filterbank to perform local normalization in the spectral domain. LNCCs were found to improve the recognition accuracy particularly when reverberation was severe. A different approach, based on the *blind spectral weighting* (BSW) technique, was proposed in [5] to handle the reverberation mismatch. In addition to these two previous methods, there is a family of reverberation-robust features based on smoothing of subband Hilbert envelopes. For example, *mean Hilbert envelope coefficient* (MHEC) feature extraction scheme, proposed in [6], takes advantage of low-pass filtering of Hilbert envelopes of Gammatone filterbank outputs. Smoothing of Hilbert envelopes can also be conducted using *frequency domain linear prediction* (FDLP) [7, 8, 9], a method to compute all-pole estimates for Hilbert envelopes. FDLP processing can be used on its own [8] or in conjunction with *time domain linear prediction* (TDLP) [10]. The technique where FDLP is followed by TDLP, known as *2-dimensional autoregressive model (2DAR)*, has been reported to provide better speaker verification results in reverberant conditions than when using FDLP alone [10]. Besides being efficient in tackling the reverberation mismatch problem, 2DAR processing has been found to improve verification in the presence of background noise as well [10].

In this study, we propose to modify the 2DAR model by replacing TDLP with *time-varying linear prediction* (TVLP) [11, 12] which is a generalization of conventional *linear prediction* (LP) [13]. TVLP can be used to analyze non-stationarity of speech signals by allowing the underlying all-pole model to be time-varying. In TVLP, temporal trajectories of the all-pole filter coefficients are represented with basis functions such as polynomials or trigonometric functions. This introduces an additional temporal constraint to 2DAR models that we hypothesize to be useful in tackling reverberation mismatch between training and testing.

Our contributions are as follows: First, we present a modification of TVLP that enables us to apply TVLP in the autocorrelation domain. Traditionally, TVLP is applied in the time-domain but to use TVLP after FDLP, it is necessary to modify the model to be applicable for spectro-temporal representations. Second, we modify the 2DAR model by replacing TDLP with TVLP. Third, we conduct speaker verification experiments with the recent text-dependent RedDots corpus to compare the proposed features with the 2DAR and MFCC features. Finally, we study the effect of RASTA filtering [14] combined with the 2DAR model, which was not included in [10].

This study was partly funded by Academy of Finland projects #284671 and #288558.

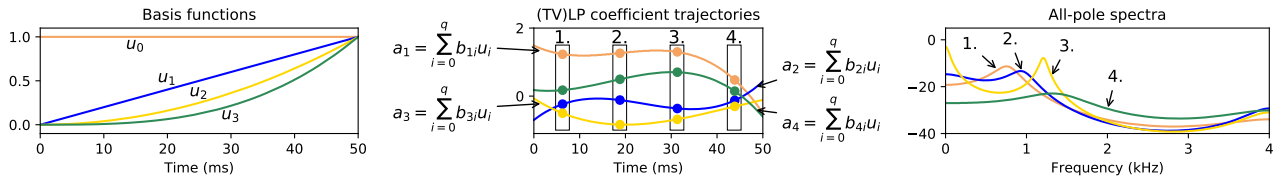

Figure 1: *An example of time-varying linear predictive (TVLP) modeling. The graph on the left shows the monomial basis functions used in TVLP-modeling of the predictor filter coefficient trajectories of 50 ms long speech signal depicted in the middle graph. In this illustration, prediction order of 4 is used, leading to only four trajectories. Graph on the right shows examples of four all-pole spectra that are all obtained from within the single 50 ms long TVLP window.*

2. Time-varying linear prediction

2.1. Classical time-domain formulation

In conventional LP analysis [13], the current speech sample $x[n]$ is predicted as a linear weighted sum of the past p samples given by $\hat{x}[n] = -\sum_{k=1}^{p} a_k x[n-k]$ where $\{a_k\}_{k=1}^{p}$ are the predictor coefficients. The solution to this formulation can be obtained by minimizing the cost function $E = \sum_n e^2[n]$, where $e[n] = x[n] - \hat{x}[n]$, which in turn leads to solving a set of normal equations given by

$$\sum_{k=1}^{p} a_k r_{ki} = -r_{0i}, \qquad i = 1, \ldots, p, \tag{1}$$

where r_{ki} denotes the *correlation coefficients* given by

$$r_{ki} = \sum_n x[n-k]x[n-i]. \tag{2}$$

The above formulation provides a piecewise constant approximation of the vocal tract system over short time intervals (frames) by assuming the system to be quasi-stationary. However, the vocal tract is a slowly but continuously varying system and therefore better modeled using *time-varying* linear prediction (TVLP) analysis over longer time intervals. TVLP introduces a time-continuity constraint on the predictor filter coefficients by expressing the prediction as

$$\hat{x}[n] = -\sum_{k=1}^{p} a_k[n]x[n-k] \tag{3}$$

where $a_k[n]$ denotes the k^{th} time-varying filter coefficient that is approximated as a linear combination of $q+1$ basis functions $u_i[n]$ as follows:

$$a_k[n] = \sum_{i=0}^{q} b_{ki}u_i[n]. \tag{4}$$

Different sets of basis functions such as monomials, trigonometric functions, or Legendre polynomials can be used for the approximation. In this paper, a simple monomial basis $u_i[n] = n^i$, $i = 0, \ldots, 3$, is adopted. An illustration of TVLP analysis using monomial basis functions is given in Figure 1.

Minimization of the cost function with respect to each polynomial coefficient leads to a set of normal equations given by

$$\sum_{k=1}^{p}\sum_{i=0}^{q} b_{ki}c_{ij}[k,l] = -c_{0j}[0,l] \tag{5}$$

for $1 \le l \le p$ and $0 \le j \le q$ [11]. Here $c_{ij}[k,l]$ denotes the *generalized correlation coefficients* defined as

$$c_{ij}[k,l] = \sum_n u_i[n]u_j[n]x[n-k]x[n-l]. \tag{6}$$

2.2. Proposed autocorrelation domain formulation

In several applications, including robust feature extraction, the signal may have been converted into a spectro-temporal representation using a filter-bank or spectrogram analysis. In such a scenario, conventional TVLP modeling of the processed signal would require a reconstruction of the time-domain signal from the spectro-temporal representation. In order to avoid any such reconstruction requiring careful handling of phase information, we propose a new *autocorrelation domain time-varying linear prediction* (AD-TVLP) analysis.

Any given spectro-temporal representation $X(t,f)$ can be converted into a sequence of autocorrelation functions $r_\tau(t) = \int_f X(t,f)\exp(-j2\pi f\tau)df$ by computing the inverse Fourier transform of the power spectrum $X(t,f)$ at each time instant t. Now, the conventional LP analysis can be applied independently on correlation function at each time instant by solving the normal equations similar to that in Eq. (1). However, a time-continuity constraint can be imposed on the LP coefficients derived at each time instant by modifying the normal equations in Eq. (1) to take advantage of the availability of a sequence of autocorrelation functions. The resulting normal equations with the continuity constraint is given by

$$\sum_{k=1}^{p} a_k[n]r_{ki}[n] \approx -r_{0i}[n], \qquad \begin{matrix} i = 1, \ldots, p, \\ n = 0, \ldots, N-1, \end{matrix} \tag{7}$$

where $r_{ki}[n]$, $a_k[n]$, and N denote the time-varying autocorrelation coefficients, the time-varying LP coefficients, and the window length for the time-varying analysis, respectively.

Substituting Eq. (4) into Eq. (7), we can compute the least squares solution to the resulting set of linear equations (of the form $\boldsymbol{Rb} = -\boldsymbol{r}$) as follows:

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} ||\boldsymbol{r} + \boldsymbol{Rb}||_2^2 \tag{8}$$

where

$$\boldsymbol{r} = [r_{01}[0], \ldots, r_{0p}[0], \ldots, r_{01}[N-1], \ldots, r_{0p}[N-1]]^T_{Np\times 1} \tag{9}$$

$$\boldsymbol{b} = [b_{10}, \ldots, b_{1q}, \ldots, b_{p0}, \ldots, b_{pq}]^T_{p(q+1)\times 1} \tag{10}$$

$$\boldsymbol{R} = [R_0, R_1, \ldots, R_{N-1}]^T_{Np\times p(q+1)} \tag{11}$$

where R_n is a $p(q+1)\times p$ matrix whose i^{th} column is given by

$$R_{ni} = \boldsymbol{r}_i[n] \otimes \boldsymbol{u}[n]. \tag{12}$$

Here, $\otimes$ denotes the Kronecker product of $\boldsymbol{r}_i[n] = [r_{1i}[n] \ldots r_{pi}[n]]^T$ and $\boldsymbol{u}[n] = [u_0[n] \ldots u_q[n]]^T$.

3. 2-D autoregressive models

3.1. Background

Two dimensional autoregressive speech modeling (2DAR) was first introduced in 2004 [15]. The 2DAR model provides a refreshing way of building speech spectrograms: instead of applying Fourier transform or AR modeling to short-time windows, the speech signal is first transformed into frequency domain and then AR modeling is applied to frequency windows followed by the usual temporal AR modeling. This idea was first adopted to speaker recognition in [16] and extended later in [17] and [10]. These studies indicate that 2DAR-processed features give consistent and considerable improvements over standard MFCC features for speaker verification in noisy conditions without compromising performance in clean conditions.

Autoregressive modeling is also known as LP modeling and hence its applications in frequency and time domain are known as *frequency domain linear prediction* (FDLP) and *time domain linear prediction* (TDLP), respectively. The former is less known, but nonetheless, it is the key concept behind the 2DAR model allowing temporal processing of speech without first splitting the signal into short-time windows.

The left side of Figure 2 shows the steps for 2DAR-processing of speech. The first step is to transform the input speech with the discrete cosine transform (DCT) and then to window this DCT-transformed signal into many overlapping frequency bands (we used 100 bands). Then, FDLP is applied to obtain all-pole estimates of Hilbert envelopes of each subband. These envelopes represent signal's energy in the subband-specific frequency range as a function of time, which allows us to form a spectrogram of the speech by stacking the information from all of the envelopes. At this stage, we perform energy integrations over the subband envelopes using 25 ms long Hamming windows at 10 ms intervals. By doing so, the spectrogram is effectively subsampled to obtain a frame rate that is similar to that used in the conventional MFCC feature extraction. This ends the FDLP part of 2DAR processing, where the data is processed along the temporal dimension.

The second part of 2DAR is to apply TDLP to the FDLP-processed spectrogram. The autocorrelation coefficients needed for computing the LP coefficients are obtained from the power spectra by using inverse Fourier transform. As a result of successive application of FLDP and TVLP, we obtain a 2DAR spectrogram that has been processed in both temporal and spectral dimensions and from which we can then extract MFCC features in the usual way.

3.2. Proposed method

We propose a modification to the 2DAR model by replacing TDLP with the autocorrelation domain TVLP. This will, in addition to spectral processing, add an extra constraint for the LP-coefficients in the temporal domain, preventing them to change too abruptly from frame to frame. This effect is demonstrated in Figure 3.

Figure 2 shows the differences between 2DAR and its modified version 2DAR-TVLP. In 2DAR-TVLP, after obtaining the autocorrelation sequences, we proceed by forming "superframes" of autocorrelation sequences by using an 11 frames long window that is shifted one frame at a time. We then apply autocorrelation domain TVLP to each of the superframes. This gives us 11 spectra per superframe and because superframes are shifted 1 autocorrelation sequence at a time, we must select only one spectrum from each superframe to keep the frame rate at the original rate (100 Hz). Thus, we extract MFCCs only from the middle frame of each superframe.

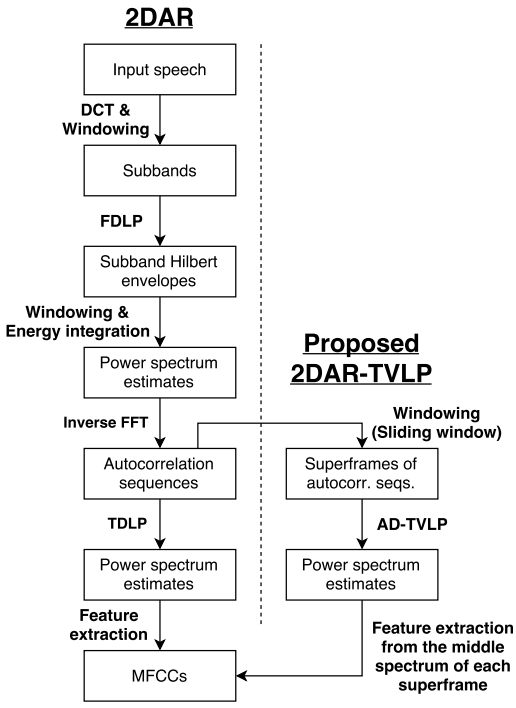

Figure 2: *Diagram showing the differences between the 2DAR and the 2DAR-TVLP models.*

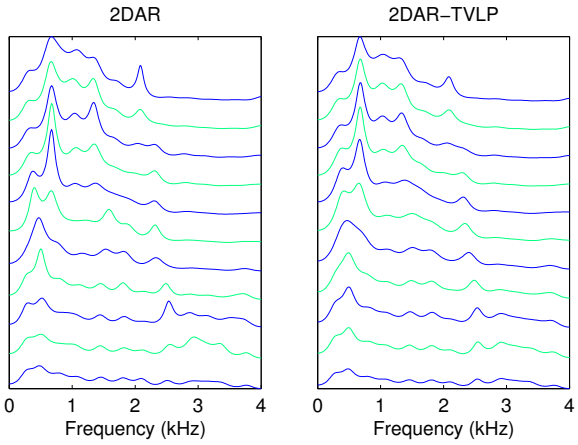

Figure 3: *11 consecutive magnitude spectra (10 ms step) obtained from 2DAR and 2DAR-TVLP processing. TVLP leads to smoother transitions between successive spectra.*

4. Experimental setup

4.1. Speech corpus

We performed speaker verification experiments using the male speakers of common phrase task of the RedDots challenge corpus [18] following the protocol for text-dependent speaker verification. This task consists of 320 pass-phrase specific target speaker models from 35 unique speakers. For enrollment, data of the same pass-phrase from three different sessions are used. The target and non-target speakers utter the same pass-phrase during enrollment and verification. The average duration of each pass-phrase is three seconds. The common phrase task has total 3242 genuine and 120086 impostor trials. As a background data, we used male speakers from TIMIT. Both corpora have a sample rate of 16 kHz.

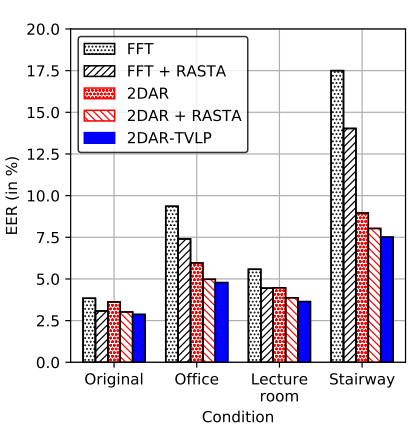

Figure 4: *Speaker verification results for different features in different conditions.*

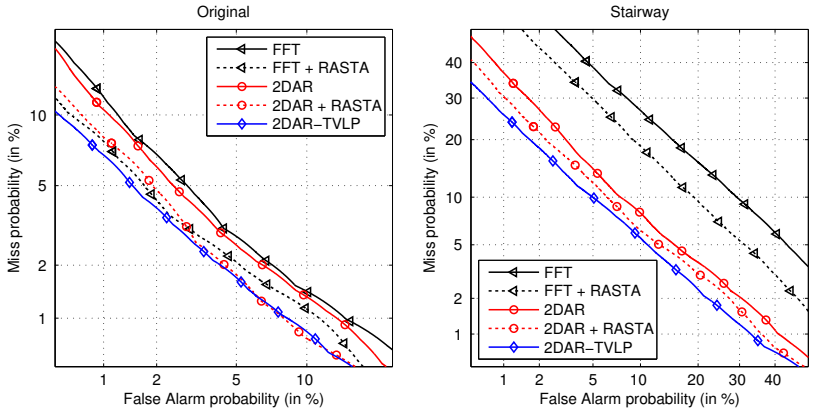

Figure 5: *Detection error tradeoff graphs for the original and stairway-reverberation conditions.*

4.2. Speech reverberation

In addition to the original RedDots data, we experimented on artificially reverberated RedDots data. Reverberation was performed by convolving original speech files with room impulse responses (RIRs) obtained from the Aachen impulse response (AIR) database [19]. We selected three different RIRs, one measured form an office room, second from a lecture room, and the last one from a stairway. Reverberation times (RT60) for these impulse responses are 0.35s, 0.28s, and 0.81s, respectively. Reverberation was only applied to the test data and not to the enrollment or background data.

4.3. Features and classifier

In our speaker verification experiments, we used the MFCC features with a configuration shown in Table 1. These features were computed from spectrograms that were obtained either by Fourier-transforming Hamming-windowed frames or by applying 2DAR or 2DAR-TVLP.

We found that minimum EER was achieved with 2DAR by using prediction orders of 24 and 42 for FDLP and TDLP, respectively. For 2DAR-TVLP, the corresponding optimal prediction orders were 24 and 38. Here, FDLP prediction orders are given for 1 second long segments and they are normalized according to the segment length. Long utterances were split into 3 second long segments before FDLP processing.

We used a classic *Gaussian mixture model – universal background model* (GMM-UBM) system [20], for which we trained a 256-component UBM from the background data. Speaker models were obtained by MAP adapting component means of UBM using relevance factor 3. We chose a lightweight GMM-UBM-based system to conduct rapid parameter experimentation with computationally heavy 2DAR models. As demonstrated in [21], GMM-UBM provides a competitive accuracy on the RedDots data consisting of short utterances.

Table 1: *Configuration of MFCC features.*

Frame length / step	25 ms / 10 ms
Number of cepstral coefficients	19
Center frequency of the first mel-filter	100 Hz
Center frequency of the last mel-filter	5400 Hz
Energy coefficient	Not included
Velocity and acceleration coefficients	Included
RASTA filter pole position (if applied)	$z = 0.97$
Speech activity detection	Energy-based [22]
Feature normalization	Cepstral mean and variance norm. (CMVN)

5. Results

Figure 4 presents the results for the speaker verification experiments on RedDots corpus in terms of EER. We compared performances of traditional FFT-based MFCCs, 2DAR-processed MFCCs, and 2DAR-TVLP-processed MFCCs. Additionally, we studied the effect of cepstral level RASTA filtering on these feature types. In preliminary experiments, we found that RASTA improves performance of FFT and 2DAR features, but does not provide benefit with 2DAR-TVLP features and hence the last combination is omitted from the figure. 2DAR-TVLP outperforms other features in all the tested conditions. Differences between 2DAR-TVLP and 2DAR with RASTA are relatively small but nevertheless consistent. Differences between 2DAR-TVLP and FFT with RASTA are larger and especially so in the reverberant conditions. Table 2 contains the exact numbers for these comparisons for the original and Stairway-reverberation conditions. Detection error tradeoff graphs in Figure 5 reveal that the good performance of 2DAR-TVLP is not restricted only to the proximity of the EER-point.

Table 2: *Speaker verification equal error rates (EER (%)) for the best performing feature configurations in the original and in the stairway-reverberation conditions. The last two columns show relative changes obtained by using the proposed features.*

	(1) FFT + RASTA	**(2) 2DAR + RASTA**	**(3) 2DAR-TVLP**	**(1)→(3) change (%)**	**(2)→(3) change (%)**
Original	3.08	3.02	2.87	-6.8	-5.0
Stairway	14.03	8.03	7.51	-46.5	-6.5

6. Conclusions

We studied the possibility of incorporating time-varying autoregressive modeling to the 2DAR feature extraction scheme to improve speaker verification in reverberant conditions. This required us to develop a new time-varying linear prediction formulation, AD-TVLP, that is applicable to the spectro-temporal representations of signals. We adopted this formulation to the existing 2DAR feature extraction method and obtained promising results. In comparison to the baseline 2DAR and MFCC features, the proposed 2DAR-TVLP features improved speaker verification performance in both original and reverberated test conditions. These promising results encourage us to further explore adopting time-varying autoregressive models for speech feature extraction in adverse conditions.

7. References

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

[2] H. Kuttruff, *Room Acoustics*, 5th ed. Spon Press, 2009.

[3] A. R. Avila, M. O. S. Paja, F. J. Fraga, D. D. O'Shaughnessy, and T. H. Falk, "Improving the performance of far-field speaker verification using multi-condition training: the case of GMM-UBM and i-vector systems." in *INTERSPEECH*, 2014, pp. 1096–1100.

[4] V. Poblete, J. P. Escudero, J. Fredes, J. Novoa, R. M. Stern, S. King, and N. B. Yoma, "The use of locally normalized cepstral coefficients (LNCC) to improve speaker recognition accuracy in highly reverberant rooms," *Interspeech 2016*, pp. 2373–2377, 2016.

[5] S. O. Sadjadi and J. H. Hansen, "Blind spectral weighting for robust speaker identification under reverberation mismatch," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 5, pp. 937–945, 2014.

[6] ——, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5448–5451.

[7] M. Athineos and D. P. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5237–5245, 2007.

[8] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *Audio Engineering Society Convention 101*. Audio Engineering Society, 1996.

[9] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1912–1924, 1999.

[10] 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 (TASLP)*, vol. 22, no. 8, pp. 1285–1295, 2014.

[11] 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.

[12] D. Rudoy, T. F. Quatieri, and P. J. Wolfe, "Time-varying autoregressions in speech: Detection theory and applications," *IEEE Transactions on audio, Speech, and Language processing*, vol. 19, no. 4, pp. 977–989, 2011.

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

[14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[15] M. Athineos, H. Hermansky, and D. Ellis, "PLP^2: Autoregressive modeling of auditory-like 2-D spectro-temporal patterns," in *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*, 2004.

[16] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-D autoregressive models for speaker recognition." in *Odyssey*, 2012, pp. 229–235.

[17] S. H. R. Mallidi, S. Ganapathy, and H. Hermansky, "Robust speaker recognition using spectro-temporal autoregressive models." in *INTERSPEECH*, 2013, pp. 3689–3693.

[18] "Reddots project," https://sites.google.com/site/thereddotsproject/home, accessed: 2017-03-15.

[19] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *Proceedings of the 16th International Conference on Digital Signal Processing (DSP)*, July 2009.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[21] H. Delgado, M. Todisco, M. Sahidullah, A. K. Sarkar, N. Evans, T. Kinnunen, and Z.-H. Tan, "Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification," in *IEEE Spoken Language Technology (SLT) Workshop*, 2016.

[22] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.