Noise-robust Linear Prediction Cepstral Features for Network Speech Recognition

Aadel Alatwi[1], Stephen So[1], Kuldip K. Paliwal[1]

[1]School of Engineering, Griffith University, Brisbane, QLD 4111, Australia

aadel.alatwi@griffithuni.edu.au, s.so@griffith.edu.au, k.paliwal@griffith.edu.au

Abstract

In this paper, we propose a perceptually-motivated method for modifying the speech power spectrum to obtain a set of linear prediction coding (LPC) parameters that possess good noise-robustness properties in network speech recognition. Speech recognition experiments were performed to compare the accuracy obtained from MFCC features extracted from AMR-coded speech that use these modified LPC parameters, as well as from LPCCs extracted from AMR bitstream parameters. The results show that when using the proposed LP analysis method, the recognition performance was on average 1.2% - 6.1% better than when using the conventional LP method, depending on the recognition task.

Index Terms: Linear prediction coding parameters; Network speech recognition; Automatic speech recognition

1. Introduction

Speech processing technologies are increasingly being incorporated into modern devices and applications such as Automatic Speech Recognition (ASR). This is partly because it can be simply used to provide service accessibility for clients. Many ASR applications are based on the Network Speech Recognition (NSR) approach, in what is known as a client-server model [1]. In this model, speech signals are compressed using conventional speech coders such as the GSM speech coder and transmitted to the server side, where feature extraction and the speech recognition are conducted [1]. There are two models of NSR systems: speech-based NSR (as shown in Figure 1), where the feature extraction is carried out on the reconstructed speech; and bitstream-based NSR (as shown in Figure 2), where the LPC parameters from the bitstream are converted to ASR features.

In the speech coder at the client side, the autocorrelation method [2] is typically used as the linear prediction coding (LPC) analysis technique to obtain the LP coefficients from short frames of speech, which are converted to some suitable LPC parameters, such as Log Area Ratios (LARs) or Line Spectral Frequencies (LSFs) [3]. These LP coefficients represent the power spectral envelope, which offers a concise representation of important properties of the speech signal. In noise-free environments, the performance of this LPC analysis technique is often satisfactory. However, in the presence of noise, the autocorrelation method yields a poor estimate of an all-pole model of the input speech signal [4]. This behavior results in an overall deterioration in the reconstructed speech quality, which also degrades the recognition performance at the server end [5].

This paper presents a new perceptually-inspired method of estimating LP coefficients, which we call the Smoothed and Thresholded Power Spectrum linear prediction (STPS-LP) coefficients. This method involves computing autocorrelation coefficients from a modified speech power spectrum, which are used in the autocorrelation method [2]. These LP coefficients can then be converted to LPC parameters that are compatible with current speech coders, with the added benefit of enabling noise-robust ASR features to be extracted on the server side. We have evaluated the effectiveness of the proposed method in comparison with conventional ASR features in terms of the recognition performance using both the speech-based and bitstream-based NSR approaches under clean and noisy conditions.

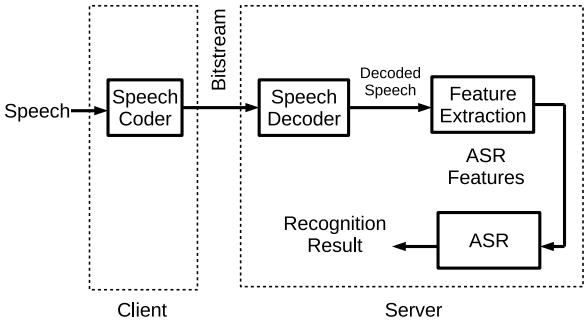

Figure 1: *Block diagram of speech-based network speech recognition (NSR).*

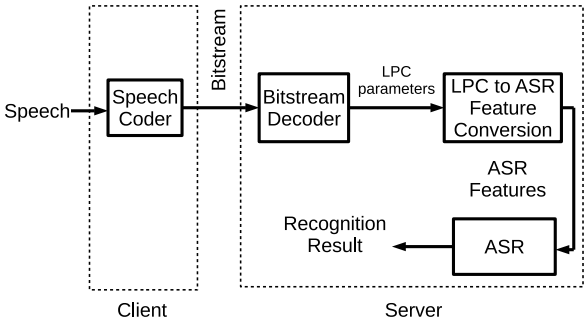

Figure 2: *Block diagram of bitstream-based network speech recognition (NSR).*

The structure of this paper is organized as follows: Section 2 explains the theory behind the proposed STPS-LP analysis method, describes the proposed algorithm, and presents the STPS-LP cepstral features at the server side. Section 3 shows the experimentally obtained results, in which we evaluate the ASR performance. Finally, we provide our conclusion in Section 4.

2. Proposed STPS-LP features for ASR

2.1. Conventional LPC analysis method

The power spectrum of a short frame $\{x(n), n = 0, 1, 2, ..., N-1\}$ of N samples of the speech signal can be modeled using an all-pole or autoregressive (AR) model [6]:

$$\hat{X}(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{1}$$

where p is the order of the AR model, $\{a_k,\ 1 \leqslant k \leqslant p\}$ are the AR parameters, and G is a gain factor. The parameters $\{a_k\}$ and G are estimated by solving the Yule-Walker equations [7]:

$$\sum_{k=1}^{p} a_k R(j-k) = -R(j), \quad \text{for } k = 1, 2, \ldots, p \tag{2}$$

$$G^2 = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{3}$$

where $R(k)$ are the autocorrelation coefficients, which are estimated using the following formula [7]:

$$R(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} x(n)x(n+k) \tag{4}$$

It can be readily shown that this AR modelling procedure of solving the Yule-Walker equations is equivalent to the autocorrelation method in linear prediction analysis [6]. In the linear prediction context, the AR parameters $\{a_k\}$ are the LP coefficients, and G^2 is the minimum squared prediction error.

The autocorrelation coefficients used in the Yule-Walker equations can also be calculated by taking the inverse discrete-time Fourier transform of the periodogram $P(\omega)$ estimate of the power spectrum [7]:

$$R(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) \mathrm{e}^{j\omega k} d\omega \tag{5}$$

where

$$P(\omega) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x(n) \mathrm{e}^{-j\omega n} \right|^2 \tag{6}$$

This provides a way of introducing preprocessing of the periodogram to reduce the variance and improve the noise robustness prior to the computation of the LP coefficients.

2.2. Estimating perceptually motivated LPC parameters

The proposed method computes the LPC parameters in two steps: In the first step, it manipulates the periodogram estimate of the power spectrum of the speech signal with the aim of reducing the variance of the spectral estimate and removing the parts that are more affected by noise. In the second step, the autocorrelation coefficients are obtained from the processed power spectrum. The processed power spectrum is obtained through smoothing followed by thresholding operations. In the smoothing operation, as shown in Figure 3, the variance of the spectral estimate is reduced by smoothing the periodogram of the input speech signal [7] using triangular filters, which are spaced using the Bark frequency scale [8]. This non-linear smoothing operation, which is inspired by the human auditory system, results in less smoothing at low frequencies, where the high power components are located, while more smoothing is applied at the higher frequencies, where weaker spectral components are more affected by noise [9]. Following the smoothing, a thresholding operation is performed, where the influence of low signal-to-noise ratio (SNR) spectral components, which are prone to being corrupted by noise and also add unnecessary variance to the spectral estimate, are removed and replaced by the smoothed spectrum, as shown in Figure 4. As a consequence of smoothing followed by thresholding, the dominant spectral peaks are preserved because they are the least affected by noise, while the less reliable spectral valleys are discarded and replaced by a smoothed average. Hence, by improving the robustness of the power spectrum estimation, the linear prediction coefficients derived from it would have lower variance and possess better robustness in noisy environments.

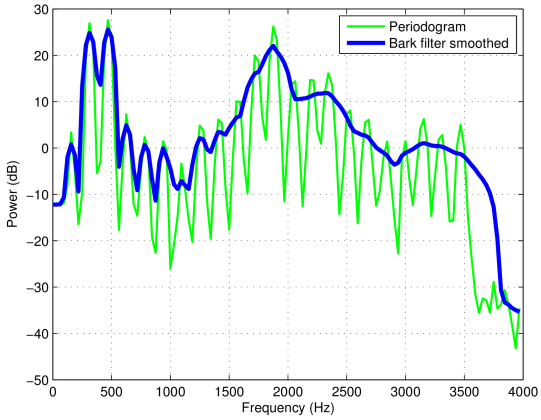

Figure 3: *Periodogram $P(k)$ and the smoothed spectrum $\bar{P}(k)$ of speech sound (vowel /e/ produced by male speaker).*

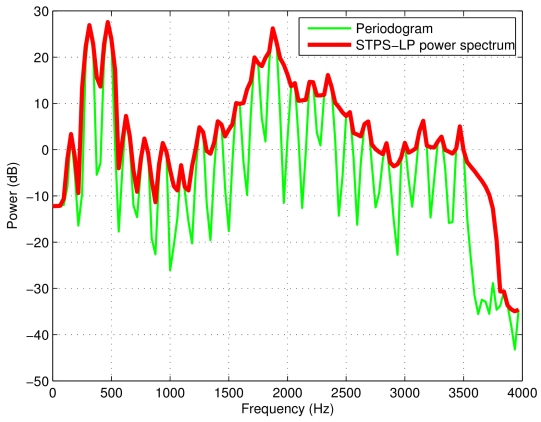

Figure 4: *Periodogram $P(k)$ and the resultant spectrum $\hat{P}(k)$ after thresholding operation of speech sound (vowel /e/ produced by male speaker).*

The proposed algorithm is described in the following steps:

Step 1: Compute the periodogram spectrum $P(k)$ of a given frame $\{x(n), n = 0, 1, 2, ..., N-1\}$ of N samples from a

speech signal [7]:

$$P(k) = \frac{1}{N}\left|\sum_{n=0}^{M-1} x(n)w(n)\mathrm{e}^{-j2\pi kn/M}\right|^2, \quad 0 \leqslant k \leqslant M-1 \tag{7}$$

where $P(k)$ is the value of the estimated power spectrum at the k^{th} normalized frequency bin, M is the FFT size where $M > N$, and $w(n)$ is a Hamming window.

Step 2: Smooth the estimated power spectrum $P(k)$ using a triangular filter at every frequency sample:

$$\tilde{P}(k) = \sum_{l=-L(k)}^{L(k)} K(l)P(l-k) \tag{8}$$

where $\tilde{P}(k)$ is the smoothed $P(k)$, $K(l)$ is the triangular filter, and $L(k)$ is half the critical bandwidth of the triangular filter at frequency sample k. The triangular filter $K(l)$ is spaced using the Bark frequency scale, which is given by [8]:

$$Bark(f) = 13\arctan(0.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \tag{9}$$

Step 3: Using the smoothed $\tilde{P}(k)$ as the threshold, $\hat{P}(k)$ is formed by retaining only spectral components that are above the threshold. This is defined as:

$$\hat{P}(k) = \begin{cases} P(k) & \text{if } P(k) \geq \tilde{P}(k) \\ \tilde{P}(k) & \text{otherwise} \end{cases}$$

Step 4: Compute the modified autocorrelation coefficients by taking an inverse discrete Fourier transform [7]:

$$\hat{R}(q) = \frac{1}{M}\sum_{k=0}^{M-1} \hat{P}(k)\mathrm{e}^{j2\pi kq/M} \quad , \quad 0 \leqslant q \leqslant M-1 \tag{10}$$

These autocorrelation coefficients $\hat{R}(q)$, $0 \leqslant q \leqslant p$, where p is the LPC analysis order, are then used in the Levinson-Durbin algorithm [7] to compute the linear prediction coefficients, which we call the Smoothed and Thresholded Power Spectrum linear prediction (STPS-LP) coefficients.

2.3. Cepstral features derived from STPS-LP coefficients for noise-robust speech recognition

For automatic speech recognition at the server end, the STPS-LP coefficients are extracted from the speech coding bitstream and then converted to a set of robust ASR cepstral-based feature vectors. In comparison with conventional LP cepstral coefficients (LPCCs), where the entire power spectrum is modeled by linear prediction analysis on a linear frequency scale, STPS-LP cepstral coefficients (or STPS-LPCCs) have the distinct advantage of being derived from a power spectrum that has been smoothed by an auditory filterbank and thresholded to remove low SNR spectral components. These operations reduce the influence of unreliable spectral components, which improve the feature's robustness to noise. We propose the following steps in the computation:

Step 1: Given the STPS-LP coefficients $\{a_k, k = 1,2,3,...,p\}$ and the excitation energy G^2, the power spectral estimate $P(\omega)$ is computed as follows [7]:

$$P(\omega) = \frac{G^2}{\left|1 + \sum_{k=1}^{p} a_k \mathrm{e}^{-j\omega k}\right|^2} \tag{11}$$

Step 2: Sample the power spectral estimate $P(\omega)$ at multiples of 0.5 bark scale, from 0.5 to 17.5 bark (to cover the range of 4 kHz), to give power spectral samples $\{\tilde{P}(r); r = 1,2,...,35\}$, where r is the sample number.

Step 3: Take the logarithm of each power spectral sample and compute the discrete cosine transform to produce a set of STPS-LPCCs [10]:

$$C(k) = \sqrt{\tfrac{2}{R}}\sum_{r=1}^{R} \log\tilde{P}(r)\cos\left[\tfrac{\pi}{R}\left(r-\tfrac{1}{2}\right)k\right], \quad 1 \leqslant k \leqslant N_c \tag{12}$$

where R = 35 and N_c is the desired number of cepstral coefficients.

3. Results and Discussion

In this section, a series of ASR tests were conducted to evaluate the NSR performance on speech that has been coded using LPC parameters that were derived from the conventional LP and STPS-LP coefficients in clean and noisy conditions. These ASR experiments were performed using MFCC (Mel frequency cepstral coefficients) features that were computed from the reconstructed speech (speech-based NSR); as well as using LPCC and STPS-LPCC features computed from the GSM coder parameters themselves (bitstream-based NSR). We utilized the Adaptive-Multi Rate coder (AMR) in 12.2 kbit/s mode, which is identical to the GSM Enhanced Full Rate [11]. There are three conditions that we tested:

- Baseline: training and testing on uncoded speech
- Matched: training on coded speech, testing on coded speech
- Mismatched: training on uncoded speech, testing on coded speech

In this study, all of the experiments were conducted using the DARPA Resource Management Continuous Speech Database (RM1) [12] under clean and noisy conditions. The training and test sets consisted of 3977 and 300 sentences, respectively. In all cases, the speech signal was downsampled to 8 kHz. For noisy conditions, the speech signal was corrupted by additive zero-mean Gaussian white noise at six different SNRs, ranging from 30 dB to 5 dB in 5 dB steps. The HTK toolkit [10] was used for the Hidden Markov Model (HMM) construction. The cepstral feature vector was composed of a 12 dimension base feature concatenated with delta and acceleration coefficients. Thus, the size of the feature vector was 36 coefficients. The recognition performance is represented by numerical values of word-level accuracy.

3.1. Recognition accuracy in speech-based NSR

Table 1 compares the performance of speech recognition accuracy using Mel Frequency Cepstral Coefficients (MFCCs) computed from the original speech signal without AMR coding (Column 2) and with AMR processed speech (Columns 3 - 6). The AMR speech was coded using the LPC parameters that were derived from the conventional LP and the proposed STPS-LP coefficients. Columns 3 and 4 show the results for

Table 1: *Word-level accuracies (%) obtained using Mel-Frequency Cepstral Coefficients derived from the original waveform and from the reconstructed speech.*

Noise Level (dB)	Baseline	Matched Models		Mismatched Models	
		LP	STPS-LP	LP	STPS-LP
Clean	95.47	94.16	94.55	93.53	93.85
30	94.46	93.57	93.65	92.12	93.18
25	92.79	92.08	92.32	91.26	91.65
20	89.58	89.07	89.27	86.17	86.99
15	77.59	72.96	75.55	68.77	71.75
10	50.31	46.11	48.53	42.40	44.79
5	16.61	14.15	15.95	12.85	13.95
Average between 5 - 30 dB	70.22	67.99	69.22	65.60	67.06

the matched condition, where the training model was computed from AMR coded speech. Columns 5 and 6 show the results for the mismatched condition, where the training model was computed from the original uncoded speech. The results show that the performance of speech-based NSR was slightly better in matched compared to mismatched models under clean condition where the speech was reconstructed using the conventional LP coefficients. This behavior did not hold in the environments of noise, especially for SNRs below 20 dB, where the performance has deteriorated in both models. On the contrary, when considering the coded speech that used the proposed STPS-LP coefficients, the provided performance is close to the baseline under almost all environmental conditions in the matched model. The recognition accuracy has improved by about 2.98% at 15 dB in comparison to AMR coding that employed the conventional method for mismatched training.

3.2. Recognition accuracy in bitstream-based NSR

Cepstral features were obtained from unquantized as well as quantized LSFs (which were derived from conventional LP and STPS-LP coefficients) that were encoded in the AMR coding bitstream. The LSF parameters (based on the conventional LP analysis method) were transformed into LP coefficients [3], and cepstral coefficients were computed using the approach described in [13] to obtain LPCCs. The proposed method that was described in Section 2.3 was used to compute STPS-LPCCs. Table 2 compares speech recognition accuracies obtained when using linear prediction cepstral features in matched conditions (training and testing based on quantized LSFs) and mismatched conditions (training based on unquantized LSFs and testing based on quantized LSFs). The results in Table 2 illustrate that, under clean conditions, there was modest improvement in the bitstream-based NSR accuracy obtained using STPS-LPCC features over LPCC features in matched and mismatched models. The STPS-LPCC features were superior to the conventional method when the speech was corrupted by white noise (that is SNR $<$ 20 dB), and in these cases the NSR performance was on an average 5.47% and 7.74% better than the conventional LPCCs in matched and mismatched models, respectively, while the baseline STPS-LPCCs was an average 6.91% better than the baseline LPCCs.

4. Conclusion

This paper has presented a new method of estimating LP coefficients that are designed to exploit the non-linear spectral selectivity of the human auditory system. These LP coefficients and their associated LPC parameters are fully compatible with industry-standard LP-based speech coders. Through smoothing and thresholding, low energy spectral components that are more vulnerable to being corrupted by noise are discarded, resulting in lower estimation variance and subsequently improved noise robustness in ASR. The speech recognition performance of the STPS-LP coefficients, in comparison with conventional LP coefficients, was investigated for two NSR scenarios. The recognition accuracy improved slightly when using MFCC features derived from speech that was coded using STPS-LP-based parameters for all SNRs and in both matched and mismatched conditions. In the bitstream NSR scenario, STPS-LPCC features computed from the bitstream parameters resulted in higher recognition accuracies, especially at lower SNRs. The results demonstrated the improved noise-robustness of the STPS-LP coefficients.

Table 2: *Word-level accuracies (%) obtained using Cepstral Coefficients derived from the LSF parameters that were transformed into the corresponding LPC coefficients.*

Noise Level (dB)	Baseline		Matched Models		Mismatched Models	
	LPCCs	STPS-LPCCs	LPCCs	STPS-LPCCs	LPCCs	STPS-LPCCs
Clean	91.98	93.08	90.34	91.75	88.74	90.77
30	88.62	90.07	87.49	89.90	84.51	87.33
25	86.16	88.35	85.02	87.25	79.66	85.33
20	83.26	86.66	81.11	84.08	76.07	80.95
15	75.28	81.07	73.72	78.14	64.84	73.29
10	57.88	66.60	56.14	62.42	47.91	56.55
5	30.07	36.30	29.48	35.18	24.64	30.78

5. References

[1] S. So and K. K. Paliwal, "Scalable distributed speech recognition using gaussian mixture model-based block quantisation," *Speech communication*, vol. 48, no. 6, pp. 746–758, 2006.

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

[3] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. New York, NY, USA: Elsevier Science Inc., 1995.

[4] S. M. Kay, "The effects of noise on the autoregressive spectral estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 5, pp. 478–485, 1979.

[5] A. Trabelsi, F. Boyer, Y. Savaria, and M. Boukadoum, "Improving lpc analysis of speech in additive noise," in *Circuits and Systems, 2007. NEWCAS 2007. IEEE Northeast Workshop on*. IEEE, 2007, pp. 93–96.

[6] J. Makhoul, "Spectral linear prediction: properties and applications," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 3, pp. 283–296, 1975.

[7] M. H. Hayes, *Statistical digital signal processing and modeling*. John Wiley & Sons, 2009.

[8] H. Fletcher, "Auditory patterns," *Reviews of modern physics*, vol. 12, no. 1, p. 47, 1940.

[9] B. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 1997.

[10] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[11] ETSI, *ETSI TS 126 090 Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); AMR speech Codec;Transcoding Functions (3GPP TS 26.090 version 7.0.0 Release 7)*. Tech. Rep., 2007.

[12] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The darpa speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, Feb 1986, pp. 93–99.

[13] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993.