Perceptually Motivated Linear Prediction Cepstral Features for Network Speech Recognition

Aadel Alatwi, Stephen So, Kuldip K. Paliwal
Signal Processing Laboratory
Griffith University, Brisbane, QLD, 4111, Australia.
Email: aadel.alatwi@griffithuni.edu.au, s.so@griffith.edu.au, k.paliwal@griffith.edu.au

***Abstract*—In this paper, we propose a new method for modifying the power spectrum of input speech to obtain a set of perceptually motivated Linear Prediction (LP) parameters that provide noise-robustness to Automatic Speech Recognition (ASR) features. Experiments were performed to compare the recognition accuracy obtained from Perceptual Linear Prediction-Cepstral Coefficients (PLP-LPCCs) and cepstral features derived from the conventional Linear Prediction Coding (LPC) parameters with that obtained from the proposed method. The results show that, using the proposed approach, the speech recognition performance was on average 4.93% to 7.09% and 3% to 5.71% better than the conventional method and the PLP-LPCCs, respectively, depending on the recognition task.**

***Index Terms*—Linear prediction coefficients; Network speech recognition; Spectral estimation**

I. Introduction

Most of the modern applications and devices using Automatic Speech Recognition (ASR) have incorporated speech processing technologies. This technology is widely utilized due to the accessibility benefits it provides to customers [1]. Many ASR devices employ Network Speech Recognition (NSR) which is known as the client-server model [2]. In the client-server approach, speech signals are compressed and transmitted to the server side using conventional speech coders such as the GSM speech coder. At the server side, the feature extraction and speech recognition are conducted [3]. There are two NSR models: speech-based network speech recognition (as shown in Fig. 1), in which the speech extraction occurs on the reconstructed speech, and bitstream-based network speech recognition model (as shown in Fig. 2), in which the linear prediction coding (LPC) parameters are converted to ASR features for speech recognition [2].

At the client side, the autocorrelation method is typically used as the LPC analysis technique to obtain the LP coefficients [4]. These LP coefficients are generated using short frames of speech, and they are then converted to suitable LPC parameters such as Log-Area-Ratios (LARs) and Line Spectral Frequencies (LSFs) [5]. The LP coefficients represent the power spectral envelope that provides a concise representation of important properties of the speech signal. In noise-free environments, the LPC analysis technique performance is highly satisfactory. However, when the noise is introduced to the environment, the results from the autocorrelation method are unreliable due to poor estimation of the all-pole spectral model of the input speech [6]. This behavior results in a severe decline in the quality of the coded speech, which further deteriorates the recognition performance at the server side [7].

This paper demonstrates the estimation of the LP coefficients using a perceptually-inspired method, attained from the Smoothed Power Spectrum Linear Prediction (SPS-LP) coefficients. In the SPS-LP method, autocorrelation coefficients are computed from a modified speech power spectrum, which are then utilized in the autocorrelation method [4]. The attained LP coefficients are then converted into LPC parameters that are well-matched with existing speech coders, with the additional advantage of allowing noise-robust ASR features to be extracted on the server side. The paper also evaluates the efficiency of the proposed approach as compared to the conventional ASR features with regard to the recognition outcome using both the bitstream-based and speech-based NSR methodologies under clean and noisy conditions.

The organization of this paper is as follows: Section II explains the theory behind the proposed approach, describes the SPS-LP algorithm, and presents the SPS-LP cepstral features at the server side. Section III shows the results from the experiments evaluating ASR. Section IV provides a conclusion of this study.

II. Proposed SPS-LP Features for ASR

A. Conventional LPC Analysis Method

The power spectrum of a short frame, represented as $\{x(n), n = 0, 1, 2, ..., N-1\}$ of N samples of the input speech signal, can be modeled using an all-pole or

978-1-5090-0941-1/16/$31.00 ©2016 IEEE

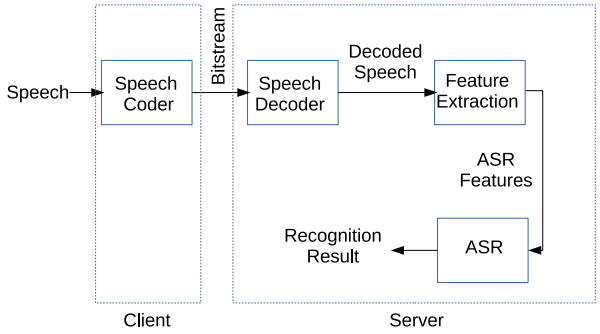

Fig. 1. Block diagram of speech-based network speech recognition (NSR).

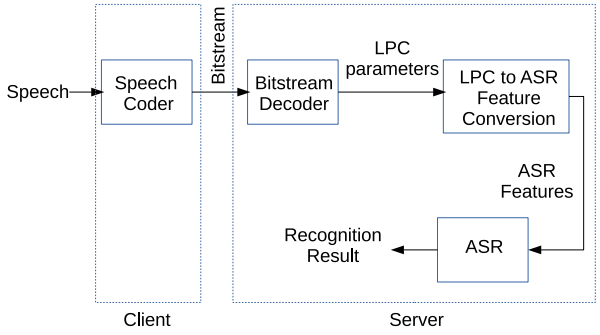

Fig. 2. Block diagram of bitstream-based network speech recognition (NSR).

autoregressive (AR) model [8]:

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

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 [9]:

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

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

where $R(k)$ are the autocorrelation coefficients, which are estimated using the following formula [9]:

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

It can be demonstrated that this AR modelling method of solving the Yule-Walker equations is equivalent to the autocorrelation method in linear prediction analysis [8]. 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 computed by taking the inverse discrete-time Fourier transform of the periodogram $P(\omega)$ estimate of the power spectrum [9]:

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

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 $P(\omega)$, which reduces the variance and improves the noise robustness prior to computation of the LP coefficients.

B. 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 input speech signal with the objective of reducing the variance of the spectral estimate and removing the parts that are more influenced by noise. In the second step, the autocorrelation coefficients are generated from the processed power spectrum. The processed power spectrum is obtained using a smoothing operation. In this smoothing procedure, as shown in Fig. 3, the spectral estimate variance is reduced by smoothing the periodogram of the input speech signal [9] using triangular filters, which are spaced using the Bark frequency scale [10]. It is well known that there is generally a downward spectral tilt in the speech power spectrum, where the higher power components tend to be located in the low frequency regions and weaker spectral components in the high frequency regions, which are more affected by noise [11] [12]. Since the effect of noise spectral components is less pronounced in the presence of high energy peaks, the non-linear smoothing process, which is inspired by the human auditory system, results in less smoothing at low frequencies and more smoothing at high frequencies. 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.

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 [9]:

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

where $P(k)$ is the value of the estimated power spectrum at the k^{th} normalized frequency bin,

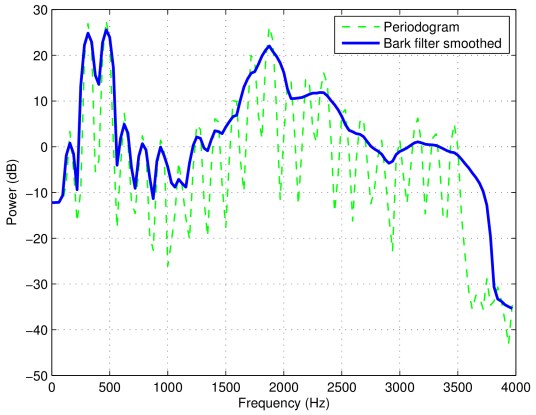

Fig. 3. Periodogram $P(k)$ and the smoothed spectrum $\bar{P}(k)$ of speech sound (vowel /e/ produced by male speaker).

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:

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

where $\bar{P}(k)$ 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 [10]:

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

Step 3: Compute the modified autocorrelation coefficients by taking the inverse discrete Fourier transform [9]:

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

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 [9] to compute the linear prediction coefficients, which we call the smoothed power spectrum linear prediction (SPS-LP) coefficients.

C. Cepstral Features Derived from SPS-LP Coefficients for Noise-Robust Speech Recognition

For automatic speech recognition at the server side, the SPS-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 power spectrum is modeled by linear prediction analysis on a linear frequency scale, SPS-LP cepstral coefficients (or SPS-LPCCs) have the distinct advantage of being derived from a power spectrum that has been smoothed by auditory filterbanks. This operation reduces the influence of unreliable spectral components, which improves the feature's robustness to noise. We propose the following steps in the computation:

Step 1: Given the SPS-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 [9]:

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

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 SPS-LPCCs [13]:

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

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

III. Results and Discussion

In this section, a sequence of ASR investigations were conducted to evaluate the NSR performance in two scenarios. In the bitstream-based NSR scenario, LPCC and SPS-LPCC features were computed from the GSM coder parameters. In the speech-based NSR scenario, PLP-LPCCs and SPS-LPCCs were generated from the reconstructed speech. All ASR investigations were carried out in clean and noisy conditions. We utilized the Adaptive-Multi Rate coder (AMR) in 12.2 kbit/s mode, which is identical to the GSM Enhanced Full Rate [14]. We tested three conditions:

- 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 (RM1) database [15] under clean and noisy conditions. 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 signal to noise ratios (SNRs), ranging from 30 dB to 5 dB in 5 dB steps. The HTK toolkit [13] was used

TABLE I
WORD-LEVEL ACCURACY (%) OBTAINED USING CEPSTRAL COEFFICIENTS DERIVED FROM THE LSFS PARAMETERS THAT TRANSFORMED INTO THE CORRESPONDING LPC COEFFICIENTS.

Feature vector		Signal to noise ratio (dB)						
		Clean	30	25	20	15	10	5
Baseline	LPCCs	91.98	88.62	86.16	83.26	75.28	57.88	30.07
	SPS-LPCCs	**92.80**	**89.70**	**87.95**	**85.66**	**80.70**	**65.45**	**35.30**
Matched Models	LPCCs	90.34	87.49	85.02	81.11	73.72	56.14	29.48
	SPS-LPCCs	**91.75**	**89.09**	**86.52**	**83.89**	**77.48**	**61.84**	**34.81**
Mismatched Models	LPCCs	88.74	84.51	79.66	76.07	64.84	47.91	24.64
	SPS-LPCCs	**89.87**	**86.65**	**84.77**	**80.95**	**72.92**	**55.86**	**29.89**

TABLE II
WORD-LEVEL ACCURACIES (%) OBTAINED USING PLP-LP AND SPS-LP CEPSTRAL COEFFICIENTS DERIVED FROM THE ORIGINAL WAVEFORM AND FROM THE RECONSTRUCTED SPEECH.

Feature vector		Signal to noise ratio (dB)						
		Clean	30	25	20	15	10	5
Baseline	PLP-LPCCs	**93.84**	89.95	88.18	85.16	78.24	59.46	33.23
	SPS-LPCCs	93.48	**90.58**	**89.65**	**86.49**	**81.72**	**67.53**	**38.82**
Matched Models	PLP-LPCCs	**92.54**	88.28	87.12	83.35	76.52	57.64	31.76
	SPS-LPCCs	92.14	**89.93**	**88.14**	**85.68**	**78.80**	**63.93**	**35.18**
Mismatched Models	PLP-LPCCs	**90.93**	86.14	84.41	79.07	71.81	54.35	27.13
	SPS-LPCCs	90.73	**87.51**	**86.02**	**81.88**	**73.83**	**57.20**	**31.25**

to construct the Hidden Markov Model. The cepstral feature vector was composed of a 12 dimension base feature including delta and acceleration coefficients. Thus, the size of the feature vector was 36 coefficients. Hence, the shape of the short-time power spectrum is used as the information that given to the recognizer, the zeroth coefficient was not included [16]. The recognition performance is represented by numerical values of word-level accuracy.

A. Recognition Accuracy in Bitstream-Based NSR

Cepstral features were obtained from unquantized and quantized LSFs (which were derived from conventional LP and SPS-LP coefficients) encoded in the AMR coding bitstream. The LSF parameters (based on the conventional LP analysis method) were transformed into the corresponding LP coefficients [5], and cepstral coefficients were generated using the approach described in [17] to obtain LPCCs. The proposed method that was described in Section *C* was used to compute SPS-LPCCs. The recognition accuracies are shown in Table I. For white noise and the level of SNR, the best score is shown in boldface. The first row of the table shows the results for the baseline condition, where the training and testing are based on unquantized LSFs, the second row shows the results for the matched condition, where the training and testing are based on quantized LSFs, and the third row shows the results for the mismatched condition, where the training is based on unquantized LSFs and testing is based on quantized LSFs.

The results indicate that, under clean conditions, there was modest improvement in the bitstream-based NSR accuracy obtained using SPS-LPCC features over LPCC features in all conditions. The SPS-LPCC features were superior to the conventional method when the speech was corrupted by white noise (SNR $<$ 20 dB), and in these cases the NSR performance was on average 4.93% and 7.09% better than the conventional LPCCs in matched and mismatched models, respectively, while the baseline SPS-LPCCs was on an average 6.07% better than the baseline LPCCs.

B. Recognition Accuracy in Speech-Based NSR

Table II illustrates the performance of speech recognition accuracy using both PLP-LPCCs and SPS-LPCCs that were computed from the original speech signal without AMR coding (Baseline) and with AMR processed speech (Matched and Mismatched Models). The PLP-LPCs were created by performing perceptual processing [18] on the AMR speech that was coded using the LPC parameters derived from the conventional LP. After this processing, we performed cepstral conversion to obtain PLP-LPCCs [17]. The SPS-LPCCs were generated from the speech that was reconstructed using the SPS-LP coefficients. In these experiments, the LP order of all-pole model was 12. The second row of the table shows the results for the matched condition, where the training model was computed from AMR coded speech. The third row of the table shows the results for the mismatched condition, where the training model was computed from the original uncoded speech. The results indicate that the performance of speech-based NSR using PLP-LPCCs was marginally improved in all models compared to SPS-LPCCs under clean condition. This behavior did not hold in the environments of noise, especially for SNRs below 20 dB, where the performance was deteriorated. On the contrary, when considering the proposed STS-LPCC features, the average recognition accuracy was improved by 5.71%, 3.99% and 3% for the baseline, matched and mismatched models, respectively.

IV. CONCLUSION

A new method of estimating LP coefficients has been presented in this paper. The proposed method was designed to exploit the non-linear spectral selectivity of the human hearing (acoustic) system. The LP coefficients and the associated LPC parameters are fully compatible with the industry-standard LP-based speech coders. Using a smoothing operation, the low energy spectral components that are more susceptible to being corrupted by noise are ignored, resulting in lower estimation variance and consequently improved noise robustness in ASR. The performance of the SPS-LP coefficients, in association with conventional LP coefficients, was investigated for the bitstream-based NSR scenario. In this scenario, SPS-LPCC features computed from the bitstream parameters resulted in higher recognition accuracies. Another comparison was performed for speech-based NSR between PLP-LPCC

and SPS-LPCC features. In this comparison, the features were computed for each method from the original and reconstructed speech. The speech recognition performance was improved especially at lower SNRs. The results demonstrate the improved noise-robustness of the SPS-LP coefficients.

REFERENCES

[1] I. Kiss, "A comparison of distributed and network speech recognition for mobile communication systems," *Proc. Int. Conf. Spoken Language Processing*, apr 2000.

[2] 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.

[3] Z.-H. Tan and B. Lindberg, *Mobile Multimedia Processing: Fundamentals, Methods, and Applications*, X. Jiang, M. Y. Ma, and C. W. Chen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

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

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

[6] 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.

[7] 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.

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

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

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

[11] P. R. Rao, *Communication Systems*. Tata McGraw-Hill Education, 2013.

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

[13] 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.

[14] 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.

[15] 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.

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

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

[18] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr 1990.