3 BRADFORD, W.J., and GRIFFITHS, H.D.: 'Digital generation of high time-bandwidth product linear FM waveforms for radar altimeters', *IEE Proc. F*, 1992, **139**, (2), pp. 160–169

4 LEWIS, B.L.: 'Range-time-sidelobe reduction technique for FM-derived polyphase PC codes', *IEEE Trans.*, 1993, **AES-29**, (3), pp. 834–840

Robust SBR method for adverse Mandarin speech recognition

Wei-Tyng Hong and Sin-Horng Chen

An RNN-based robust signal bias removal (RRSBR) method is proposed for improving both the recognition performance and the computational efficiency of the SBR method for adverse Mandarin speech recognition. It differs from the SBR method in using three broad-class sub-codebooks to encode the feature vector of each frame and combining the three encoding residuals to form the frame-level signal bias estimate. A novel approach involving softly combining the board-class encoding residuals using dynamic weighting functions generated by an RNN is applied. Experimental results show that the RRSBR method significantly outperforms the SBR method.

Introduction: Recently, many studies have been devoted to the field of robust speech recognition in adverse environments corrupted with background noise and channel distortion. Among them, the signal bias removal (SBR) method [1] has been shown to be a promising method for overcoming channel distortion. The SBR method assumes that the effect of channel distortion is to add a bias to the original speech signal in the cepstrum domain, and hence uses a two-step iterative procedure to remove the signal bias. The first step involves estimating the signal bias by calculating the average encoding residual of the testing utterance using a pre-trained codebook. The second step involves subtracting the bias estimate from every frame of the testing utterance. If all bias-corrupted input frames are encoded to the proper codewords to which their clean-speech counterparts are encoded, the average encoding residual will be an unbiased estimate of the channel bias. This makes the iterative procedure converge very quickly. But this is only a theoretical case. In practical situations, some input frames will be encoded to improper codewords to interfere with the signal bias estimation. The iterative procedure will suffer from degraded efficiency and may even lead to convergence to a value which deviates significantly from the real channel bias. The situation is more serious when the channel distortion is large or when both channel distortion and background noise exist simultaneously. To overcome this drawback, a recurrent neural network (RNN)-based robust SBR (RRSBR) method for adverse Mandarin speech recognition is proposed for improving both the efficiency and the accuracy of signal bias estimation in the SBR method.

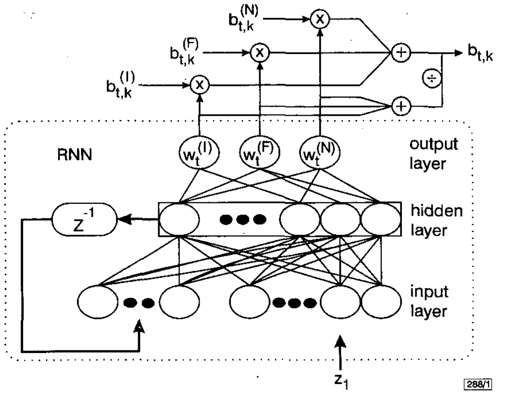

Fig. 1 *Schematic diagram of RNN-based encoding residual combination approach*

Proposed RRSBR method: The RRSBR method uses a similar two-step iterative procedure to the conventional SBR method to estimate and remove the signal bias. The only difference lies in the method of calculating the frame-level signal bias in the first step. Instead of using a single large codebook which covers the entire signal space, the RRSBR method uses three smaller sub-codebooks to encode the feature vector of the current frame and then combines the encoding residuals to form the frame-level signal bias estimate. These three sub-codebooks represent three broad classes of signals including *syllable initial, syllable final*, and non-speech. A novel approach involving softly combining the three encoding residuals using dynamic weighting functions generated by an RNN is proposed here. Fig. 1 shows a schematic diagram of the RNN-based residual combination approach. It can be seen from Fig. 1 that the RNN is a three-layer network with all outputs of the hidden layer being fed back to itself as additional inputs. It operates in a frame-synchronous mode to generate three dynamic weighting functions for the three broad classes of *initial, final*, and non-speech [2, 3]. It is noted that the first two broad classes are chosen because of the simple *initial-final* structure of Mandarin base-syllables. For other languages, they can be changed to the two broad classes of unvoiced and voiced sounds. The three dynamic weighting functions, $\{w_t^{(I)}, w_t^{(F)}, w_t^{(N)}\}$, are used to softly combine the three encoding residuals, $\{b_{t,k}^{(I)}, b_{t,k}^{(F)}, b_{t,k}^{(N)}\}$ by

$$b_{t,k} = \left(\sum_{\forall c} w_t^{(c)} b_{t,k}^{(c)}\right)\left(\sum_{\forall c} w_t^{(c)}\right)^{-1} \qquad (1)$$

where the subscripts t and k denote the frame and iteration indexes, $c \in \{I, F, N\}$ is the broad-class index,

$$b_{t,k}^{(c)} = z_{t,k-1} - \rho_{j_t^{(c)}}^{(c)} \qquad (2)$$

represent the encoding residuals for broad-class c at the kth iteration, $\rho_{j_t^{(c)}}^{(c)}$ is the nearest neighbour codeword in sub-codebook c,

$$\begin{aligned} j_t^{(c)} &= \arg\max_j \Pr\left(z_{t,k-1} | \rho_j^{(c)}, \Sigma_j^{(c)}\right) \\ &= \arg\max_j N\left(z_{t,k-1}, \rho_j^{(c)}, \Sigma_j^{(c)}\right) \qquad (3) \end{aligned}$$

is the index of the nearest neighbour codeword, $z_{t,k-1}$ is the bias-removed observed signal at frame t of the (k–1)th iteration, and $N(\cdot, \rho, \Sigma)$ denotes a multivariate normal distribution with mean vector ρ and (diagonal) covariance matrix Σ. The signal bias estimate $\hat{b}_k$ for the whole utterance at the kth iteration is then calculated by

$$\hat{b}_k = \frac{1}{T}\sum_{t=1}^{T} b_{t,k} \qquad (4)$$

where T is the length of the utterance. In the second step, the bias-removed observed signal of the kth iteration is calculated by

$$z_{t,k} = z_{t,k-1} - \hat{b}_k \quad \text{for } t = 1, ..., T \qquad (5)$$

By iteratively performing the above two steps, a local optimal estimate of the channel distortion can be obtained. The iterative procedure can be initialised by subtracting an initial bias estimate $\hat{b}_0 = \bar{z} - \bar{\mu}$ from all frames of the testing utterance, where $\bar{z}$ is the cepstrum mean of the utterance and $\bar{\mu}$ is the cepstrum mean of the entire training database [4].

Table 1: Recognition results of RRSBR and SBR methods

	Number of iterations	Clean	42dB	30dB	18dB
		%	%	%	%
SBR	1	59.4	56.1	48.4	23.1
	10	65.1	64.1	52.1	26.3
RRSBR	1	74.7	72.0	57.3	29.4
	10	75.0	72.5	58.3	31.3

Experimental results: The effectiveness of the RRSBR method was examined on a multispeaker Mandarin speech recognition task using a simulated adverse-speech database generated from a clean-speech database. The clean-speech database was generated by 10 speakers and consisted of in total 3050 utterances including 2572 training utterances (12800 syllables) and 478 testing utterances (2666 syllables). To generate the adverse-speech database, each clean-speech testing utterance was first corrupted by computer-generated white Gaussian noise and then passed through a filter which simulated a telephone channel. Noise at levels of 18, 30 and 42dB in SNR was added to all utterances of the clean-speech testing database. To simulate the channel variations in the telephone speech signals through the public switching network, a set of 77 simulated filters was generated from a large telephone-speech database. It is noted that the stationarity of the environment characteristics for each utterance is guaranteed in this simulated adverse-speech database via the use of an utterance-dependent channel filter and noise level. All speech signals were first preprocessed for each 20ms Hamming-windowed frame with 10ms shift. A set of 25 recognition features including 12 MFCCs, 12 delta MFCCs, and a delta log-energy was then computed for each frame. A set of sub-syllable HMM models containing 100 three-state right-*final*-dependent *initial* models and 39 five-state context-independent *final* models was trained from the clean-speech database [5]. A single-state model was used for non-speech. Observation features in each HMM state were modelled by a mixture Gaussian distribution with diagonal covariance matrix. The number of mixture components in each state was variable and depended on the number of training samples, but a maximum value of 20 was set. The three sub-codebooks were formed by collecting all mixture components of the 100 *initial* models, of the 39 *final* models, and of the non-speech model, respectively. For performance comparison, the SBR method was also tested. The codebook used in the SBR method was formed as the union of the three sub-codebooks used in the RRSBR method. Table 1 shows the base-syllable accuracy rates of the RRSBR and SBR methods. As can be seen from Table 1, the RRSBR method outperformed the SBR method significantly. The RRSBR method also converged very rapidly. A single iteration was almost sufficient in this test. In contrast, the SBR method required many iterations to obtain a gradual performance improvement.

Summary: A new RNN-based robust SBR method for adverse Mandarin speech recognition has been presented. Experimental results have confirmed that it significantly outperforms the conventional SBR method in terms of both recognition rate and computational efficiency.

© IEE 1999 29 March 1999
Electronics Letters Online No: 19990637
DOI: 10.1049/el:19990637

Wei-Tyng Hong and Sin-Horng Chen (*Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China*)

E-mail: schen@cc.nctu.edu.tw

References

1 RAHIM, M., and JUANG, B.H.: 'Signal bias removal by maximum likelihood estimation for robust telephone speech recognition', *IEEE Trans. Speech Audio Process.*, 1996, **4**, pp. 19–30

2 HONG, W.T., and CHEN, S.H.: 'A robust RNN-based pre-classification for Noisy Mandarin speech recognition'. Eurospeech-97, 1997, Vol. 3, pp. 1083–1110

3 CHEN, S.H., LIAO, Y.F., CHIANG, S.M., and CHANG, S.: 'An RNN-based pre-classification method for fast continuous Mandarin speech recognition', *IEEE Trans. Speech Audio Process.*, 1998, **6**, pp. 86–90

4 ZHAO, Y.: 'Self-learning speaker and channel adaptation based on spectral variation source decomposition', *Speech Commun.*, 1996, **18**, pp. 65–77

5 WANG, Y.-R., and CHEN, S.-H.: 'Mandarin telephone speech recognition for automatic telephone number directory service'. ICASSP-98, 1998, Vol. 2, pp. 841–844

Short message service link for automatic vehicle location reporting

N. Papadoglou and E. Stipidis

A novel system integration using the global positioning system (GPS), short message service (SMS) and controller area network (CAN) systems for real time applications is described. These applications could be used effectively to assist automatic vehicle location (AVL) with benefits in terms of increased public security, reduced traffic congestion, and increased access to commercial information.

Introduction: The Federal Communications Commission (FCC) has introduced recent requirements whereby wireless network operators will be required to implement an E-911, the 999 UK equivalent, location capability with 125m accuracy by the year 2001 for public safety. Alternative solutions are provided by the recent integration of the global positioning system (GPS) with cellular telephony [1] and, in the automotive industry, Fieldbuses, in particular CAN systems, have already been deployed for vehicle-reporting (VR) systems where information as to the status of the vehicle is transmitted over the wireless medium.

The work introduced here relates to a novel system architecture where a short message service (SMS) can be used to periodically transmit vehicle location and control/status information with the aid of a controller area network (CAN) and GPS. The fact that an SMS is categorised as a 'virtual circuit' data service where network information is transferred without dedicating a physical channel provides an ideal link for the proposed system. Periodic vehicle location and control information can be transmitted by a large number of users without having to increase the mobile network resources. The feasibility study and preliminary results indicate that the novel system can provide a reliable real time VR service.

System architecture and functionality: The simplified integrated system architecture comprises three individual systems. The GPS is a satellite navigation system that provides location information according to the universal time co-ordinate (UTC) with an accuracy of between 25 and 100m. The CAN system that corresponds to the physical and data link layers of the ISO reference model is a Fieldbus used extensively in real time vehicle control applications. The communication link considered is the GSM-SMS where messages are transmitted in signalling channels via an SMS-Centre (SMS-C).

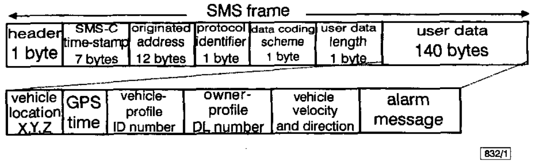

Fig. 1 *SMS frame format*

The location information extracted by the GPS receiver is placed on the CAN bus and is sent to the vehicle control (V/C) section every 1s period. The GPS follows the National Marine Electronics Association (NMEA) interface standard that allows electronic equipment to exchange navigation data in ASCII protocol NMEA-0183. The data are transmitted in the form of 'sentences' that contain up to 82 characters including a two letter 'talker ID' and a carriage return/line fee (CR/LF). On the CAN serial bus other information is transmitted such as data from oil pressure, brakes, and crash sensors where the V/C assembles this information. The CAN uses short messages with a maximum utility load of 94bits where 64bits is the maximum payload. For a connectionless type of communication, the CAN serial bus is connected directly to all the communication nodes within the system and allows transmission rates up to 1Mbit/s in real time applications. A message is created within the V/C containing the current and previous vehicle location, direction of movement, velocity, UTC transmitted time, and vehicle/user identification (Fig. 1). This message is scheduled for transmission every 5s to a reporting centre (RC) through the SMS-C where with an additional algorithm the positioning of the car can be more accurately predicted.