tion $rmx = k+s \bmod \emptyset(p)$ and signature verification $y^{rm} = r\alpha^s \bmod p$), only the legitimate signer with knowledge of $x$ can generate the signatures to satisfy the verification. Although the attacker can generate bogus signatures in the signature collection protocol, these signatures cannot satisfy the batch verification criterion. There are some other secure ElGamal type signature schemes as proposed in [4] that can also be used to design similar DSA type secure interactive batch verification protocols. We list those schemes in Table 1.

**Table 1:** Secure ElGamal type signature schemes

| | Signature equation | Signature verification |
|---|---|---|
| (1) | $mx=rk+s \bmod \emptyset(p)$ | $y^m=r^r\alpha^s \bmod \emptyset p$ |
| (2) | $sx=rk+m \bmod \emptyset(p)$ | $y^s=r^r\alpha^m \bmod \emptyset p$ |
| (3) | $sx=k+mr \bmod \emptyset(p)$ | $y^s=r\alpha^{mr} \bmod \emptyset p$ |
| (4) | $(r+m)x=k+s \bmod \emptyset(p)$ | $y^{r+m}=r\alpha^s \bmod \emptyset p$ |
| (5) | $sx=k+(m+r) \bmod \emptyset(p)$ | $y^s=r\alpha^{m+r} \bmod \emptyset p$ |

*Conclusion:* Instead of using an insecure batch verification criterion as proposed by the Naccache *et al.* in Eurocrypt '94, we propose several secure batch verification criteria in this Letter. By using the interactive batch verification protocol, the signer follows the signature collection protocol to generate $n$ signatures through interactions with the verifier and the verifier validates all these $n$ signatures at once through the batch verification criterion.


*6 December 1994*
*Electronics Letters Online No: 19950203*

L. Harn (*Computer Science Telecommunications Program, University of Missouri – Kansas City, MO 64110, USA*)

**References**

1 NACCACHE, D., M'RAIHI, D., RAPHEALI, D., and VAUDENAY, S.: 'Can DSA be improved: Complexity trade-offs with the digital signature standard'. Pre-proc. Eurocrypt'94, 1994, pp. 85–94
2 NIST: 'Digital signature standard' (FIPS PUB XX, 1993)
3 LIM, C.H., and LEE, P.J.: 'Security of interactive DSA batch verification', *Electron. Lett.*, 1994, **30**, (19), pp. 1592–1593
4 HARN, L., and XU, Y.: 'Design of generalised ElGamal type digital signature schemes based on the discrete logarithm', *Electron Lett.*, 1994, **30**, (24), pp. 2025–2026

## Speaker-independent Mandarin plosive recognition with dynamic features and multilayer perceptrons

W.-Y. Chen and S.-H. Chen

*Indexing terms: Neural networks, Speech recognition*

A new method for recognising plosives in isolated Mandarin syllables is discussed in the Letter. After automatically detecting the plosive segment of the input utterance, some dynamic features are extracted from its spectral parameter contours using orthonormal polynomial transforms. Next, an MLP trained with an algorithm based on a minimum error criterion is employed to distinguish plosives using these features. A promising recognition rate of 73.6% is achieved in a speaker-independent test using a database containing utterances of 110 syllables uttered by 100 speakers.

*Introduction:* Each character in Mandarin speech is pronounced as a syllable. An isolated Mandarin syllable can be decomposed phonetically into initial and final subsyllable units. Only 22 initials (including a dummy one) and 39 finals are available in Mandarin speech. Six of these 22 initial subsyllables are plosives, /b, d, g, p, t, k/. Similar to the vocabulary in the English E-set, Mandarin syllables with the same final subsyllable and different plosive initial subsyllables form a confusing set. Identifying these six plosives still remains the most challenging task in Mandarin speech recognition. Previous related investigations have emphasised the selection of effective features for recognition [1–3]. Wang [1, 2] *et al.* suggested extracting some features from the burst spectrum, the format transition and the voice-onset time (VOT). Next, an MLP was employed to recognise the three unaspirated Mandarin plosives /p, t, k/. Although a high recognition rate was obtained, the method was tested only on a small database containing utterances of a small vocabulary (nine syllables with /p,t,k/ followed by /i,a,u/) as generated by seven speakers. Besides, recognition features were not automatically extracted. Manual preprocessing should be required to detect the VOT.

In this study, a new method for recognising the six plosives in isolated Mandarin syllables is discussed. In this method, the plosive part of the input testing utterance is first detected automatically. For the plosive segment, the percepture linear predictive (PLP) [4] features are extracted. In the PLP analysis technique, some properties of hearing, e.g. the critical-band spectral resolution, the equal-loudness pre-emphasis, and the intensity loudness power law, are simulated to obtain an auditory-like spectrum. The PLP features are the cepstral coefficients of an autoregressive all-pole model of the auditory-like spectrum of speech. Next, these PLP features are transformed into another feature set by using orthonormal polynomial transforms to represent the dynamics of spectral parameter contours. An MLP trained with an algorithm based on the minimum error criterion [5] is then employed to recognise the plosive. Notably, the number of recognition features is fixed and independent of the length of the testing utterance. Therefore, the time alignment between the testing plosive segment and the MLP recogniser is not required. Moreover, the performance of the proposed method is tested in a speaker-independent recognition mode using a database containing 100 repetitions of utterances of 110 syllables which are all possible combinations of the six plosives and 39 final subsyllables.

*Orthonormal polynomial transform:* The speech signal is a dynamic signal in nature. As an utterance is divided into segments, each segment should be treated as a dynamic rather than a static signal. This phenomenon is especially true for plosives to recognise in this study. Therefore, instead of representing each parameter contour of a plosive segment by a constant curve, approximating it by a smooth curve is preferred. Distortion can thereby be reduced owing to the better curve fitting. In this study, the smooth curve is a reconstructed version of the original parameter contour obtained by orthonormal polynomial expansion using several low-order coefficients. Specifically, a parameter contour of a plosive segment with length $N + 1$ frames is allowed to be denoted by $f(n/N)$, $n = 0, \ldots, N$. The smooth curve used to approximate it can then be expressed by

$$\hat{f}\left(\frac{n}{N}\right) = \sum_{j=0}^{r} \alpha_j \phi_j\left(\frac{n}{N}\right) \quad 0 \le n \le N \qquad (1)$$

where

$$\alpha_j = \frac{1}{N+1} \sum_{n=0}^{N} f\left(\frac{n}{N}\right) \phi_j\left(\frac{n}{N}\right)$$

and $r$ is the order of the orthonormal polynomial expansion. As $r = 2$, the first three basis functions of the orthonormal polynomial transform can be expressed as [6]

$$\phi_0\left(\frac{n}{N}\right) = 1 \qquad (2)$$

$$\phi_1\left(\frac{n}{N}\right) = \left[\frac{12N}{N+2}\right]^{\frac{1}{2}} \left[\frac{n}{n} - \frac{1}{2}\right]$$

$$\phi_2\left(\frac{n}{N}\right) = \left[\frac{180N^3}{(N-1)(N+2)(N+3)}\right]^{\frac{1}{2}} \left[\left(\frac{n}{N}\right)^2 - \frac{n}{N} + \frac{N-1}{6N}\right]$$

for $0 \le n \le N$ and $N \ge 3$. Notably, all these three basis functions are normalised, in length, to [0,1]. After performing orthonormal polynomial transforms, coefficients of all parameter contours are accumulated and fed into the following MLP recogniser for plosive discrimination. If there are $p$ spectral parameter contours, $p(r + 1)$ recognition features are obtained.

*Speech database:* A database provided by Telecommunication Laboratories, MOTC, ROC, was used to examine the performance of the proposed plosive recognition method. The database contains utterances of 110 syllables which are all possible combinations of the six plosives and 39 final subsyllables. It was generated by 100 speakers, i.e. 50 male and 50 female. Each speaker uttered these 110 isolated syllables only once. Utterances of 76 speakers were taken for the training, and all others were used for testing. All speech signals were sampled at a 10kHz rate and digitised at 16bit resolution. After excluding a few misrecorded and mispronounced data, the number of usable tokens was recorded as 10629 (out of 110 syllables× 100 speakers = 11000).

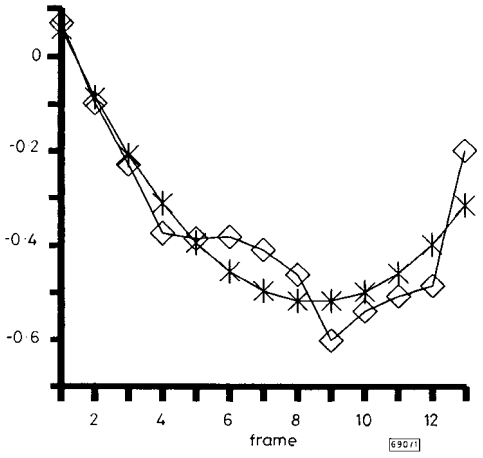


**Fig. 1** *Original output contour of cepstral coefficient CO and its reconstructed contour using second-order orthonormal polynomial expansion*

Input utterance is /k/ of /kāng/

◇ cepstral coefficient CO
✳ reconstructed contour

*Speech analysis:* The procedure of feature extraction is listed as follows:

(i) The endpoints of each utterance were detected based on the energies and zero crossing rates [7].

(ii) The detected speech signal was bandpass filtered from 1 to 3.5 kHz by a fourth-order Butterworth filter to enhance the energy discontinuity of the boundary between the plosive and the following final subsyllable. This situation occurs because much of the energy of a plosive is located beyond the frequency range 1–3.5 kHz.

(iii) With a frame length 25.6ms and a frame rate 8ms, the frame with the maximum energy difference is regarded as the plosive/final subsyllable boundary. The search is from the beginning frame to the frame having the maximum energy peak.

(iv) For each frame in the detected plosive segment, 12 cepstral coefficients are derived by the 15th-order PLP model.

(v) There are a total of 26 dynamic parameter contours (energy, 12 cepstral coefficients and their time differences) used to characterise each plosive segment.

(vi) For each parameter contour, three low-order coefficients are extracted by an orthonormal polynomial expansion. Excluding the zeroth coefficients of both the energy and the delta energy contours, there are a total of 76 recognition parameters.

**Table 1:** Confusion matrices

| | | Recognised outputs | | | | | |
|---|---|---|---|---|---|---|---|
| | | b | d | g | p | t | k |
| Unknown inputs | b | 280 | 56 | 22 | 13 | 4 | 6 |
| | d | 47 | 403 | 27 | 6 | 16 | 1 |
| | g | 16 | 47 | 347 | 6 | 3 | 10 |
| | p | 20 | 21 | 9 | 238 | 86 | 27 |
| | t | 9 | 19 | 9 | 56 | 320 | 34 |
| | k | 19 | 8 | 40 | 29 | 24 | 305 |

Recognition rate = 73.6%

*Simulation results:* Validation of the feature extraction is first investigated when testing the efficiency of the features. A typical example of some reconstructed smooth curves and the corresponding original contours are displayed in Fig. 1. This Figure reveals that all reconstructed smooth curves match quite well with their original counterparts. Hence, these polynomial expansion coefficients are efficient parameters for representing the dynamics of the parameter contours of a plosive. Next, plosive discriminations by the MLP recogniser based on these 76 features are performed. The MLP has one hidden layer comprising 35 nodes and six output nodes to represent the six Mandarin plosives. The MLP was trained by using a generalised probabilistic descent algorithm [5], which is based on a criterion to minimise the recognition error rate. In the training, the learning rate is initially set to 0.1 and would then linearly decay as the training progresses. Table 1 lists the confusion matrix of the recognition result. A promising recognition rate of 73.6% is obtained.

*Acknowledgment:* The authors thank the Telecommunication Laboratories, MOTC, ROC, for providing the database.


*19 December 1994*
*Electronics Letters Online no: 19950188*

W.-Y. Chen and S.-H. Chen (*Department of Communication Engineering, National Chiao Tung University, Hsinchu 30050, Taiwan, Republic of China*)

**References**

1 WANG, H.C., LIU, L.C., LEE, L.M., and CHANG, Y.C.: 'A study on the automatic recognition of voiceless unaspirated stops', *J. Acoust. Soc. Am.*, 1991, **89**, (1), pp. 461–464

2 LIU, L.C., LEE, L.M., WANG, H.C., and CHANG, Y.C.: 'Layered neural nets applied in the recognition of voiceless unaspirated stops', *IEE Proc. I*, 1991, **138**, (2), pp. 69–75

3 CHAN, C., and NG, K.W.: 'Separation of fricatives from aspirated plosives by means of temporal spectral variation', *IEEE Trans.*, 1985, **ASSP-33**, (4), pp. 1130 1137

4 HERMANSKY, H.: 'Perceptual linear predictive (PLP) analysis of speech', *J. Acoust. Soc. Am.*, 1990, **87**, (4), pp. 1738–1752

5 KATAGIRI, S., LEE, C.H., JUANG, B.H., JUANG, B.H., KUNG, S.Y., and KAMM, C.A.: 'New discriminative training algorithms based on the generalized probabilistic descent method'. Proc. IEEE Neural Networks for Signal Process (NSSP), 30 September – 2 October 1991, pp. 299–308

6 CHEN, S.H., and WANG, Y.R.: 'Vector quantization of pitch information in Mandarin speech', *IEEE Trans.*, 1990, **COM-38**, (9), pp. 1317–1320

7 RABINER, L.R., and SAMBUR, M.R.: 'An algorithm for determining the endpoints of isolated utterances', *Bell Syst. Technol. J.*, 1975, **54**, (2), pp. 297–315