higher resolution of time-frequency by wavelets and its capabilities for non-stationary analysis is useful for detecting the sudden burst in the signal that is useful for classification. In addition these features are found to be more robust than the features of MFCC and the 24-band filter based on AWP.

5 October 2001
Electronics Letters Online No: 20011029
DOI: 10.1049/el:20011029

O. Farooq and S. Datta (*Department of Electronic and Electrical Engineering Loughborough University, Loughborough, LE11 3TU, United Kingdom*)

References

1 TUFEKCI, Z., and GOWDY, J.N.: 'Feature extraction using discrete wavelet transform for speech recognition'. Proc. IEEE Southeastcon 2000, Nashville, USA, 2000, pp. 116–123
2 FAROOQ, O., and DATTA, S.: 'Dynamic feature extraction by wavelet analysis'. Proc. 6th Int. Conf. Spoken Language Processing, Beijing, China, Oct. 2000, Vol. 4, pp. 696–699
3 CHANG, S., KWON, Y., and YANG, S.-I.: 'Speech feature extracted from adaptive wavelet for speech recognition', *Electron. Lett.*, 1998, **34**, (23), pp. 2211–2213
4 LONG, C.J., and DATTA, S.: 'Discriminant wavelet basis construction for speech recognition'. Proc. 5th Int. Conf. Spoken Language Processing, Sydney, Australia, Nov./Dec. 1998, Vol. 3, pp. 1047–1049
5 LUKASIK, E.: 'Wavelet packets based features selection for voiceless plosives classification'. Proc. ICASSP 2000, Istanbul, Turkey, 2000, Vol. 2, pp. 689–692
6 FAROOQ, O., and DATTA, S.: 'Mel filter-like admissible wavelet packet structure for speech recognition', *IEEE Signal Process. Lett.*, 2001, **8**, (7), pp. 196–198
7 MALLAT, S.: 'A wavelet tour of signal processing' (Academic Press, 1998)

Text-independent speaker identification based on explicit exploitation of stochastic characteristics of test utterance

W.H. Tsai, W.W. Chang and C.S. Huang

The benefit of exploiting the stochastic characteristics of test utterance for speaker identification (speaker ID) via cross likelihood ratio and Bayesian information criterion is explored. Simulation results demonstrate the superiority of the proposed approaches over the conventional speaker ID based on maximum likelihood decision rule.

Introduction: Conventional speaker-identification (speaker-ID) systems [1] have been using maximum likelihood decision rule to hypothesise the identity of a test speaker by determining which client speaker model best matches the test utterance. This approach, however, does not fully compare the statistical similarities between the test speaker's voice and the voice of each client speaker. Specifically, a test utterance is simply used to compute the likelihood scores for client speaker models, while the stochastic characteristics themselves are largely ignored. As a result, such a system may suffer from unreliable likelihood scores owing to the defective models. To compensate for this shortcoming, we propose to bilaterally compare the stochastic characteristics between the test speaker's voice and the voice of the client speaker, instead of simply taking the unilateral likelihoods into account. This study investigates two approaches based on cross likelihood ratio and Bayesian information criterion, respectively.

Cross likelihood ratio: Denote $\mathbf{Y}_i = \{\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, ..., \mathbf{y}_{i,T_{y,i}}\}$ as the $T_{y,i}$-length feature vectors extracted from enrolment speech of the ith client speaker, $1 \le i \le P$. The cross likelihood ratio (CLR) [2] between $\mathbf{Y}_i$ and an unknown test utterance $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{T_x}\}$ is defined by

$$CLR(\mathbf{X}, \mathbf{Y}_i) = \log \frac{p(\mathbf{X}|\lambda_{y,i})}{p(\mathbf{Y}_i|\lambda_{y,i})} + \log \frac{p(\mathbf{Y}_i|\lambda_x)}{p(\mathbf{X}|\lambda_x)} \quad (1)$$

where λ_x and $\lambda_{y,i}$ are the stochastic models (e.g. Gaussian mixture model, (GMM)) formed from $\mathbf{X}$ and $\mathbf{Y}_i$, respectively. CLR defined above accounts not only for how well a particular client speaker model matches a test utterance, but also for how well a test speaker model matches the enrolment speech of a particular client. A larger CLR signifies a higher similarity between the test speaker's voice and the voice of a client speaker, and therefore an identifier should decide in favour of a speaker $\hat{S}$ satisfying

$$\hat{S} = \arg \max_{1 \le i \le P} CLR(\mathbf{X}, \mathbf{Y}_i) \quad (2)$$

Bayesian information criterion: The Bayesian information criterion (BIC) [3] is a model selection criterion, which aims at choosing one among a set of candidate models having different model complexities to best represent the given data. Denote $\{\Lambda_i \mid 1 \le i \le K\}$ as the candidate model set and $O = \{o_1, o_2, ..., o_T\}$ the data set, the BIC for model Λ_i is defined by

$$BIC(\Lambda_i) = \log p(O|\Lambda_i) - \frac{1}{2}\gamma d_i \log T \quad (3)$$

where $p(O|\Lambda_i)$ is the maximised model likelihood, γ is a penalty weight with value between 0 and 1, and d_i is the number of free parameters in model Λ_i. Because a higher-complexity model usually matches the data better, a penalising term is added to the log-likelihood so as to balance the measurements for non-nested models. In applying the concept of BIC to the problem of speaker ID, we consider the following two hypothesis tests, H_0 and H_1. Denote $\mathbf{Z}_i = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{T_x}, \mathbf{y}_{i,1}, \mathbf{y}_{i,2}, ..., \mathbf{y}_{i,T_{y,i}}\}$ as the concatenated sequence of $\mathbf{X}$ and $\mathbf{Y}_i$, and let λ_x, $\lambda_{y,i}$ and $\lambda_{z,i}$ be the stochastic models formed from $\mathbf{X}$, $\mathbf{Y}_i$ and $\mathbf{Z}_i$, respectively.

H_0: $\mathbf{X}$ and $\mathbf{Y}_i$ are produced by the same speaker, and thus it is appropriate to use a single model $\lambda_{z,i}$ for clarifying such a production.

H_1: $\mathbf{X}$ and $\mathbf{Y}_i$ are produced by different speakers, and therefore two separate models λ_x and $\lambda_{y,i}$ are needed to characterise the individual speaker's voice.

To evaluate which of the two hypotheses is favourable, we compute the BIC difference between H_0 and H_1:

$$\begin{aligned}\Delta BIC_i(H_0, H_1) &= BIC(\lambda_{z,i}) - BIC(\lambda_x, \lambda_{y,i}) \\ &= \log \frac{p(\mathbf{Z}|\lambda_{z,i})}{p(\mathbf{X}|\lambda_x)p(\mathbf{Y}_i|\lambda_{y,i})} \\ &\quad + \frac{1}{2}\gamma[d_x \log T_x + d_{y,i} \log T_{y,i} - d_{z,i} \log(T_x + T_{y,i})] \quad (4)\end{aligned}$$

where d_x, $d_{y,i}$ and $d_{z,i}$ denote the number of free parameters in model λ_x, $\lambda_{y,i}$ and $\lambda_{z,i}$, respectively. In general, $\Delta BIC_i(H_0, H_1)$ is large if $\mathbf{X}$ and $\mathbf{Y}_i$ are of the same speaker, and small otherwise.

In practice, computing the delta BIC above requires two extra models (λ_x and $\lambda_{z,i}$) to be trained during the testing phase, making it rather expensive when GMMs are employed. To alleviate this problem, we propose using multiple uni-Gaussian models instead of a single GMM for speaker modelling. Specifically, feature vectors of each client speaker, $\mathbf{Y}_i$ is segmented into N_i subsequence $\mathbf{Y}_{i,k}$ having length $T_{y,i,k}$, $1 \le k \le N_i$, and the delta BIC is computed for each pair of ($\mathbf{X}$, $\mathbf{Y}_{i,k}$). Assume that $\mathbf{X} \sim G(\mu_x, \Sigma_x)$, $\mathbf{Y}_{i,k} \sim G(\mu_{y,i,k}, \Sigma_{y,i,k})$, and $\mathbf{Z}_{i,k} \sim G(\mu_{z,i,k}, \Sigma_{z,i,k})$, then

$$\begin{aligned}&\Delta BIC_{i,k}(H_0, H_1) = \\ &\frac{1}{2}[(T_x + T_{y,i,k}) \log|\Sigma_{z,i,k}| - T_x \log|\Sigma_x| - T_{y,i,k} \log|\Sigma_{y,i,k}|] \\ &- \frac{1}{4}\gamma(M^2 + 3M) \log(T_x + T_{y,i,k}) \quad (5)\end{aligned}$$

where M is the dimension of the feature vectors. The decision rule takes into account the average of the delta BIC between test utterance and all subsequences of enrolment data, i.e.

$$\hat{S} = \arg \max_{1 \le i \le P} \frac{1}{N_i} \sum_{k=1}^{N_i} \Delta BIC_{i,k}(H_0, H_1) \quad (6)$$

Experimental results: Speech data used for this study consist of a subset of The 1999 NIST Speaker Evaluation Database, in which 25 males and 25 females were chosen as clients to conduct a close-set speaker-ID experiment. Two one-minute conversations, involving two different handsets, were used as enrolment data of each client. The test set consists of around 300 s conversational speech

per speaker. Speech features including 12 mel-frequency cepstral coefficients (MFCCs) and 12 delta MFCCs were extracted with 20 ms frame rate and 10 ms frame shift. Prior to training and testing, cepstral mean normalisation was applied to minimise channel-induced perturbations. The evaluation of speaker-ID experiment was performed in a segment-by-segment manner, with 1 s segment rate and 0.5 s segment shift. Each segment was treated as a separate test utterance. The speaker-ID rate was computed as the percentage of correctly identified segments over all test segments.

Within the CLR-based framework, the number of mixtures used in each of the client speaker GMMs were empirically set to be 32, while the number of mixtures used in a test speaker GMM was investigated for 1, 2 and 4, respectively. Within the BIC-based framework, feature vectors of each client speaker were empirically segmented into six uniform subsequences (each corresponds to around 20 s duration) in this experiment, and the delta BIC was computed as expressed in eqn. 5. The penalty weight γ was empirically determined to be 0.7. For performance comparison, a baseline system that uses conventional GMMs with diagonal covariance matrices was also evaluated. Fig. 1 summaries the speaker-ID results of the various approaches. As expected, the three approaches yield better performance with an increase in test utterance length. Compared with the baseline system, both the CLR-based approach and the BIC-based approach are superior to the conventional approach based on maximum likelihood decision rule, especially when the length of test utterance increases.

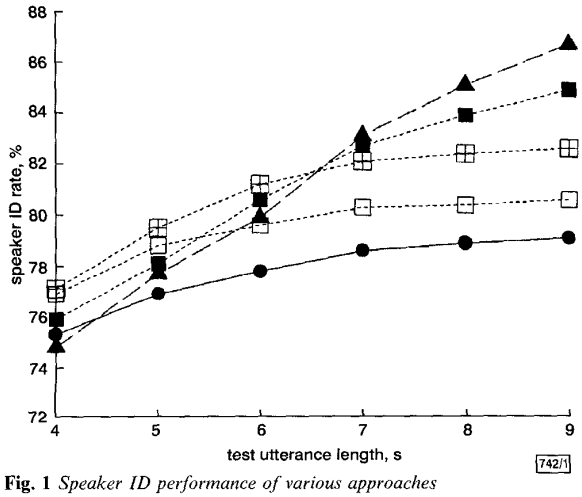

Fig. 1 *Speaker ID performance of various approaches*

—●— baseline system (32-mixture GMM/speaker)
- -□- - CLR using 1-mixture GMM for test utterance
- -⊞- - CLR using 2-mixture GMM for test utterance
- -■- - CLR using 4-mixture GMM for test utterance
– ▲ – BIC

Conclusions: Explicit exploitation of the stochastic characteristics of test utterance for speaker ID has been validated via simulations on text-independent task. It is worth noting that while this study presented only experimental results in close-set identification task, the design techniques can be applied to more general problems in speaker recognition and language identification.

17 September 2001
Electronics Letters Online No: 20011044
DOI: 10.1049/el:20011044

W.H. Tsai and C.S. Huang (*Philips Research East Asia-Taipei, Taiwan, Republic of China*)

W.W. Chang (*Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China*)

References

1 REYNOLDS, D.A., and ROSE, R.C.: 'Robust text-independent speaker identification using Gaussian mixture speaker models', *IEEE Trans. Speech Audio Process.*, 1995, **3**, (1), pp. 72–83
2 JUANG, B.H., and RABINER, L.R.: 'A probabilistic distance measure for hidden Markov models', *AT&T Tech. J.*, 1985, pp. 391–408
3 SCHWARZ, G.: 'Estimating the dimension of a model', *Ann. Stat.*, 1978, **6**, pp. 461–464

Fast IP packet delineator

R.A. Bourne and C.I. Phillips

A fast-acting synchronisation mechanism for Internet Protocol (IPv4) packet delineation from an ingress bitstream is presented. It builds on delineation and scrambling mechanisms developed for asynchronous transfer mode (ATM) technology, but provides enhancements that cater for variable-sized packets. Implications of the scheme for IP version 6 are considered.

Introduction: Typically, IP data is transported over datalink layer technologies such as asynchronous transfer mode (ATM) or point-to-point protocol (PPP) that, in turn, are carried by physical and optical layer protocols. Each of these protocols provides a number of services to the layer above. Frequently, this can lead to a duplication of functionality. Protocol de-layering is a process where several layers of a protocol stack are implemented as a single layer or eliminated, avoiding this duplication. De-layering a protocol stack can also enable various alarm timers to be set more stringently. For example, to avoid the inappropriate triggering of network layer restoration activity during the period when datalink or physical layer functions may be resolving the error condition, it is normal practice to set alarm timers higher up the stack to larger values. Removing lower layer continuity functionality allows for fast-acting IP layer protection to be implemented, possibly providing a more cost-effective solution.

This Letter describes a novel and efficient IP packet delineation mechanism that provides a vital step towards protocol de-layering by removing the need for datalink layer framing. It can be readily implemented in hardware, providing wire-speed delineation within local and metropolitan area networks.

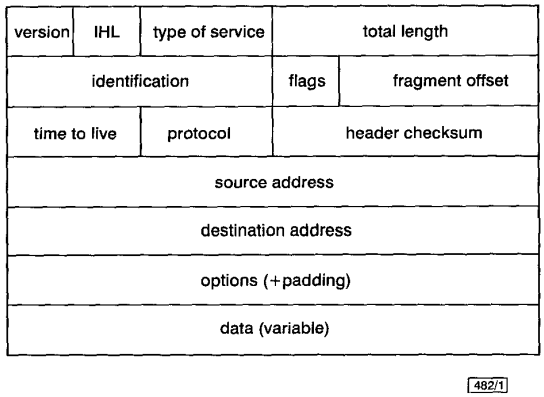

Fig. 1 *IPv4 packet format*

Operation with IPv4: A detailed definition of IP is provided in [1] and the IPv4 packet format is shown in Fig. 1. Traditionally, delineation of IP packets, i.e. the identification of the packet boundaries within a data stream, is provided by the underlying datalink layer technology that encapsulates them. This is because IP is asynchronous, uses variable length packets and variable length headers, provides no byte alignment and has no explicit frame alignment sequence.

However, IP version 4 (IPv4) provides a number of reference points within the header that are either invariant, or vary in an easily verifiable manner. For example, the first *nibble* in every IPv4 packet contains the binary version field identifier '0100'. Next, the header checksum field is always present at a fixed offset to this version field. This operates on the entire IP packet header including options fields, should they exist. The checksum algorithm is the 16 bit one's complement of the one's complement sum of all 16-bit words in the header. For the purposes of computing the checksum, the value of the checksum field is zero. This is simple to compute and can be readily achieved using combinational logic.

The packet delineation functionality is illustrated in Fig. 2. To obtain and maintain synchronisation the design assumes that the ingress stream is an abutted series of IPv4 packets. When user packets are unavailable for transmission, empty packets are inserted into the flow in a similar manner to the idle cells of the