A Novel Embedded Speaker Verification on System on Chip

Pengfei Mao
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
Beijing, China
e-mail: mpf06@mails.tsinghua.edu.cn

Jia Liu
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
Beijing, China
e-mail: liuj@tsinghua.edu.cn

***Abstract*—This paper realizes a Text-Independent, Speaker Verification system on a System on Chip (SOC) platform. The system uses Mel-Frequency Cepstral Coefficients (MFCC) features with a Gaussian Mixture Model-Universal Background Model （GMM-UBM） speaker model. To deal with resource limitations, a new speaker-centric score normalization technique is introduced. This normalization technique results in a relative EER reduction of 44.9% compared to no normalization.**

Keywords-speaker verification; system on chip; score normalization

I. INTRODUCTION

Embedded speaker recognition systems are becoming more important with the rapid development of the handheld and other portable devices. However, few products are yet available due to high chip cost. In this paper we introduce a custom SOC designed for speaker verification.

To accomplish this, a Gaussian Mixture Model-Universal Background Model（GMM-UBM）system for Text-Independent Speaker Verification is implemented on a 16-bit DSP core. The balance of cost and performance is achieved by choosing a 16-bit microcontroller with 16-bit coprocessor platform, by carefully adjust the training and recognizing algorithms, and by adding some low complexity heuristic methods. The resulting system has 128 Gaussian mixtures with diagonal covariance, with mean-only adaptation to maximize model storage efficiency.

The database used in this system is from the speech database of National 863 Program Office for Intelligent Computing Topics [1]. The content is from People's Daily. The sampling frequency is 16 kHz in a quiet recording environment.

II. HARDWARE ARCHITECTURE

The block diagram of the SOC is shown in Figure 1, and details of each block are described below.

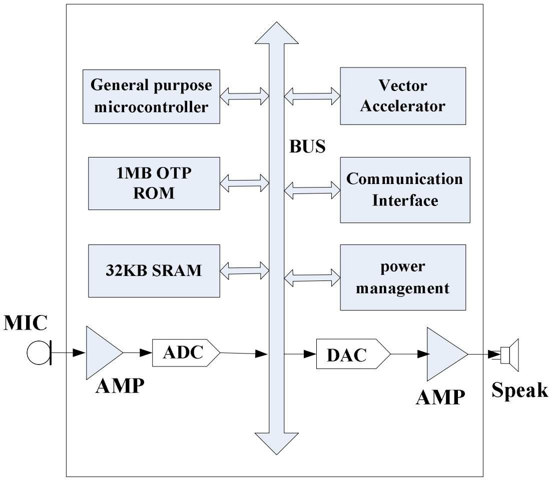

Figure 1. Block diagram of the SOC

The speaker verification system-on-chip is comprised of a general purpose microcontroller, 32KB SRAM, 1MB OTP ROM, vector accelerator, 16-bit ADC and DAC, analog filter circuit, audio input and output amplifiers, and communication interface. A power management module and a clock generator are also integrated in the SOC. In addition, the SOC integrates 3 sets of 16-bit Timers, 24 general purpose I/O ports and other peripheral circuits. The inclusion of a vector accelerator leads to a significant increase in computing power. Compared with a general purpose DSP, this ASIC integrates ADC, DAC, audio amplifier and power management modules, and omits several unnecessary circuits, so that the cost becomes much lower.

A Reduced Instruction Set Computer （RISC）8bit MCU core is used in the general purpose microcontroller. The instructions are optimized in accordance with computing characteristics of the algorithm. The internal access clock of MCU may be set in software, with a typical work clock of 60 MHz. The running frequency of the clock is quite low in power-saving mode. The SOC integrates a 32 KB RAM and 1 MB one time programmable（OTP） ROM. OTP ROM, rather than FLASH ROM, is able to reduce cost and has more flexibility than masked ROM. As a

978-0-7695-3735-1/09 $25.00 © 2009 IEEE
DOI 10.1109/FSKD.2009.714

key component for implementing high-speed computation, the vector accelerator handles 16-bit fixed point vector arithmetic with a main purpose of calculating the inner product of two vectors.[2]

III. Software algorithms

For text-independent speaker recognition, where there is no prior knowledge of what the speaker will say, the historical approach has been Gaussian Mixture Model-Universal Background Model（GMM-UBM）. The structure of the GMM-UBM system is shown in Figure 2 below.

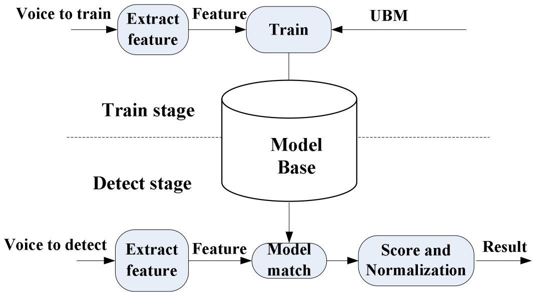

Figure 2. Structure of the GMM-UBM system

The basis for the speaker verification system is the GMM used to represent speakers, where the distribution of feature vectors extracted from a person's speech is modeled by a Gaussian mixture density. For a D-dimensional feature vector denoted as x, the mixture density for speaker s is defined as

$$p(\mathbf{x}/\lambda_s)=\sum_{i=1}^{M}p_i^s b_i^s(\mathbf{x}). \qquad (1)$$

The mixture weights p_i^s, furthermore satisfy the constraint $\sum_{i=1}^{m}p_i^s=1$. The density is a weighted linear combination of M component unimodal Gaussian densities $b_i^s(\mathbf{x})$, each parameterized by a mean vector, μ_i^s, and covariance matrix, $\sum_i^s$,

$$b_i^s(\mathbf{x})=\frac{1}{(2\pi)^{D/2}\left|\mathbf{\Sigma}_i^s\right|^{1/2}}$$

$$\times\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i^s)'\mathbf{\Sigma}_i^{s-1}(\mathbf{x}-\boldsymbol{\mu}_i^s)\right\}. \qquad (2)$$

Collectively, the parameters of speakers' density model are denoted as $\lambda_s=\left\{p_i^s,\boldsymbol{\mu}_i^s,\mathbf{\Sigma}_i^s\right\}$, $i=1,2\cdots,M$ [3].

While the general model form supports full covariance matrices, in this paper diagonal covariance matrices are used. This choice is based on empirical evidence that diagonal matrices outperform full matrices and the fact that the density modeling of an Mth order full covariance mixture can equally well be achieved using a larger order, diagonal covariance mixture.

In the GMM-UBM system we use a single, speaker-independent background model. The UBM is a large GMM trained to represent the speaker-independent distribution of features. The method to train the UBM is to merely pool all the data via the Expectation-Maximization (EM) algorithm [4]. The hypothesized speaker model is individually formed by adapting the parameters of the UBM using the speaker's training speech and MAP adaptation [5].

In this paper, we a use 27-dimensional feature vector, which contains 12 Mel-Frequency Cepstral Coefficients (MFCC), the 12-dimension MFCC first differences, the normalized energy, and its first and second order difference.

IV. Score normalization

One of the most important problems in speaker verification is to find the optimal speaker independent threshold. One approach to maintaining independence is through score normalization. Most of the current score normalization techniques' illustrated in the NIST evaluations are based on impostor models [6]. However, in a practical system, there are more true speaker voices and few impostor voices. So a speaker-centric score normalization is preferred. There are two speaker-centric score normalization introduced here.

A. Mean normalization

Test scores for the true speaker and impostor speakers tend to correlated, in that if the true speaker score is high the impostor scores will also be high. To address this, the scores can be normalized using the mean score value average access multiple test sentences. The formula is given below. In equation (3), the S_{mean} can be obtained from the N test sentence scores ($S_1,S_2,\ldots\ldots,S_N$) for a given speaker model. So the normalized scores can be calculated from equation (4).

$$S_{mean}=\frac{S_1+S_2+\ldots\ldots+S_N}{N} \qquad (3)$$

$$S_1^* = S_{\mathrm{ori}} / S_{mean} \quad (4)$$

It should also be noted that it is possible to update the normalization constant S_{mean} without reading to store all scores or directly computer by equation (3). Given a new score S_{N+1}, and the current S_{mean} and number of speakers N, an updated constant can be computed via

$$S_{mean}^* = \frac{N * S_{mean} + S_{N+1}}{N+1} \quad (5)$$

$$N^* = N + 1 \quad (6)$$

B. Distribution normalization

A common normalization technique which uses a mean and variance estimation for distribution scaling is zero normalization (Z-norm) [7]. To implement this, a speaker model is tested against example impostor utterances and the log-likelihood scores are used to estimate a speaker specific mean and variance for the impostor distribution. Zero normalization has the form below.

$$X^{'} = \frac{X - \mu_{\mathrm{I}}}{\sigma_{\mathrm{I}}} \quad (7)$$

Where $X^{'}$ is the normalized variable and μ_{I} & σ_{I} are the normalization parameter.

This normalization technique is exactly the same as the method just introduced, only with the addition of a variance form. The original log-likelihood score S_{ori} is converted using

$$\mu_d = \frac{S_1 + S_2 + \ldots\ldots + S_N}{N} \quad (8)$$

$$\sigma_d = \sqrt{\frac{(S_1 - \mu_d)^2 + (S_2 - \mu_d)^2 + \ldots\ldots + (S_N - \mu_d)^2}{N}} \quad (9)$$

$$S_2^* = \frac{S_{\mathrm{ori}} - \mu_d}{\sigma_d} \quad (10)$$

When updating the parameter (μ_d, σ_d), we must store the data ($S_1, S_2, \ldots\ldots, S_N$). With a new score S_{N+1}, we can calculate the parameter (μ_d, σ_d) in (8) and (9).

V. EVALUATION RESULT

The data used in this system is from the speech database of National 863 Program Office for Intelligent Computing Topics [1], which includes 83 males and 83 females with more than 520 sentences each. The content is from People's Daily and the sampling rate is 16 KHz, with data collected in a quiet recording environment.

The background UBM model was trained in advance using the speech from 45 men and 45 women. The remainder 38 men and women were used as test data. For each test speaker, 10 sentences were used for adaptation to create a new speaker model from the UBM model.

For each test model, one true speaker test and 14 impostor tests are conducted, with each test consisting of two sentences of data. To accomplish the normalization, 100 iterations of this process were implemented, using the results to calculate S_{mean}, μ_d and σ_d. Final EER results were calculated in a series of 40 additional true tests. Each individual test runs as described above, consists of 76 true tests (38 male and 38 female speaker) plus 1064 impostor tests (14 impostors times 76 male and female speaker), for a total of 1140 verifications. The 76 true and 1064 impostor test results are matched with a threshold to determine the EER for each test case.

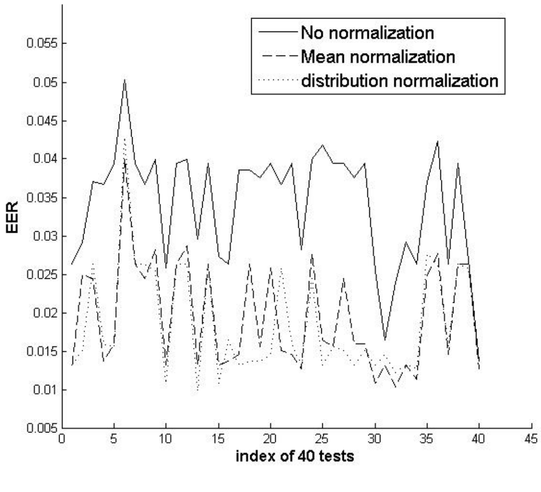

Figure 3. The result of the evaluation

The result of the evaluation is shown in Figure 3. The vertical axis is the EER and the horizontal axis is the sequence number of the 40 tests. There are three results for each test. These three results for each test are obtained by no normalization, mean normalization and distribution normalization. In 39 of the 40 tests, EER is decreased through the use of either mean normalization or the distribution normalization. The effect of these two normalizations is similar.

In mean normalization, EER is decreased by a maximum of 63.1% and 42.8% on average. In the distribution normalization, EER is decreased by a maximum of 68.5% and 44.9% on average.

The EER of the distribution normalization is less than that of the mean normalization. However when using distribution normalization, we need to store enough scores to estimate the parameters (μ_d , σ_d), requiring more memory and computational effect. Mean normalization requires less computation for similar result.

VI. Conclusion

In this paper, a speaker verification system is implemented on a SOC platform. Two new score normalization methods which are speaker-centric are introduced. The result of this system is good with the new normalization with an average EER improvement of 42.8%. As the cost of SOC is very low, it is well suited for product development. Future work will include studying the methods to reduce model complexity and reduce training time.

Acknowledgment

This project is supported by National Natural Science Foundation of China and Microsoft Research Asia (No. 60776800), and by the National High Technology Research and Development Program of China (863 Program) (No. 2006AA010101, No. 2007AA04Z223 and No. 2008AA02Z414).

References

[1] Jingfang Zhou, Research of Handset Compensation and Speaker Segmentation in Telephone Speaker Recognition, Master Thesis of Tsinghua University, 2004.

[2] Haijie Yang, Jing Yao, Jia Liu, "A Novel Speech Recognition System-on-Chip", *Audio, Language and Image Processing, 2008, ICALIP 200*8, 764-768.

[3] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing 10, 19–41 (2000),2000.

[4] Reynolds, D. A. and Rose, R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. Speech Audio Process. 3 (1995), 72–83.

[5] Gauvain, J. L. and Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Trans. Speech Audio Process. 2 (1994), 291–298.

[6] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems", Digital Signal Processing 10, 42–54 (2000).

[7] Reynolds, D., "Comparison of background normalization methods for text-independent speaker verification". In Proc. Eurospeech 1997, Rhodes, 1997, pp. 963–966.