SVM Based Speaker Verification Using Sparse Maximum A Posteriori Adaptation

Younggwan Kim, Jaeyoung Roh, and Hoirin Kim

Department of Electrical Engineering, Korea Advanced Institute of Science and Technology / Daejeon 305-701, Korea
{cleanthink, ezmalark}@kaist.ac.kr, hrkim@ee.kaist.ac.kr

***** Corresponding Author: Younggwan Kim

Received July 14, 2013; Revised July 29, 2013; Accepted August 12, 2013; Published October 31, 2013

** Short Paper*

***Abstract*:** Modern speaker verification systems based on support vector machines (SVMs) use Gaussian mixture model (GMM) supervectors as their input feature vectors, and the maximum a posteriori (MAP) adaptation is a conventional method for generating speaker-dependent GMMs by adapting a universal background model (UBM). MAP adaptation requires the appropriate amount of input utterance due to the number of model parameters to be estimated. On the other hand, with limited utterances, unreliable MAP adaptation can be performed, which causes adaptation noise even though the Bayesian priors used in the MAP adaptation smooth the movements between the UBM and speaker dependent GMMs. This paper proposes a sparse MAP adaptation method, which is known to perform well in the automatic speech recognition area. By introducing sparse MAP adaptation to the GMM-SVM-based speaker verification system, the adaptation noise can be mitigated effectively. The proposed method utilizes the L0 norm as a regularizer to induce sparsity. The experimental results on the TIMIT database showed that the sparse MAP-based GMM-SVM speaker verification system yields a 42.6% relative reduction in the equal error rate with few additional computations.

***Keywords*:** Speaker verification, GMM, SVM, MAP adaptation, Sparse representation

1. Introduction

Speaker verification systems determine if the input utterance is spoken by the claimed speaker. In recent years, a Gaussian mixture model (GMM) and support vector machine (SVM) based speaker verification systems [1-5] have provided better performance than the conventional GMM and universal background model (UBM) based systems [6]. The speaker verification is essentially a two-class problem and SVM was originally developed for two class classifier. Therefore, SVM can provide a good solution for the speaker verification task. In addition, SVM can employ a kernel function that provides nonlinear mapping between the input feature space and SVM feature space. The GMM supervector kernel [1], which is used in GMM-SVM speaker verification systems, can be thought of as a sequence kernel that maps the input features onto SVM feature space. In the conventional GMM-SVM speaker verification system, the maximum a posteriori (MAP) adaptation, which provides an accurate estimation with a moderate amount of adaptation data [7], is used to generate speaker dependent GMMs by adapting the UBM. On the other hand, the length of the input utterance may be limited in specific applications, such as entrance security systems. Although the Bayesian priors used in the MAP adaptation provide a certain smoothing operation between the model parameters of the speaker dependent GMM and UBM, small unreliable movements of the model parameters that are not relevant to the input utterance can still constitute adaptation noise.

In the automatic speech recognition area, the sparse MAP adaption method for speaker adaptation [8, 9] was proposed to mitigate the adaptation noise and reduce memory requirements. The sparse MAP adaptation employs sparsity-promoting constraints, such as the L0 or L1 norm in the MAP adaptation criteria. Through these constraints, sparse MAP adaptation can remove the small differences that constitute adaptation noise, and can consequently improve the MAP adaptation in the speech

recognition domain.

This paper uses a sparse MAP adaptation in a GMM-SVM speaker verification system to mitigate the adaptation noise that is caused by the limited amount of data. The experimental results are presented by the equal error rate (EER) and detection error tradeoff (DET) curve, and these measures show results demonstrating that the performance of the sparse MAP-SVM system is obviously higher than that of the conventional MAP-SVM system.

The remainder of this paper is structured as follows. Section 2 describes the conventional MAP-SVM speaker verification system. Section 3 introduces the sparse MAP adaptation. Finally, Sections 4 and 5 respectively compare the sparse MAP-SVM system with the conventional MAP-SVM system according to verification results and conclude the study.

2. MAP Adaptation-Based GMM-SVM Speaker Verification System

2.1 Maximum A Posteriori Adaptation

The basic approach of generating speaker dependent GMM supervectors is to derive the speaker dependent models by adapting the speaker independent model (i.e., UBM). The speaker independent model parameters were obtained using sufficient training data to train all the parameters properly. This enables tighter coupling between the speaker-dependent model and UBM, which produces better performance than the estimation based on the maximum likelihood criterion.

Assume that the UBM is given by the following:

$$p(\mathbf{x}) = \sum_{i=1}^{G} \lambda_i^{\mathrm{UBM}} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i^{\mathrm{UBM}}, \boldsymbol{\Sigma}_i^{\mathrm{UBM}}), \tag{1}$$

where $\mathcal{N}(\cdot)$ is a normal distribution, G is the number of mixtures, and λ_i^{UBM}, $\boldsymbol{\mu}_i^{\mathrm{UBM}}$, and $\boldsymbol{\Sigma}_i^{\mathrm{UBM}}$ are the weight, mean vector, and covariance matrix of each mixture, i, respectively. A diagonal covariance matrix was assumed for $\boldsymbol{\Sigma}_i^{\mathrm{UBM}}$. MAP adaptation adapting the UBM is composed of two major stages. In the first stage, sufficient statistics of the adaptation data required to calculate the mixture mean parameters were computed for each mixture of the UBM. Suppose that a set of acoustic feature vectors extracted from the utterance of the hypothesized speaker $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$ is given. The posterior probability of the mixture i given feature vector $\mathbf{x}_t$ is then calculated with all mixtures $i \in \{1, \cdots, G\}$ as follows.

$$P(i \mid \mathbf{x}_t) = \frac{\lambda_i^{\mathrm{UBM}} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_i^{\mathrm{UBM}}, \boldsymbol{\Sigma}_i^{\mathrm{UBM}})}{\sum_{g=1}^{G} \lambda_g^{\mathrm{UBM}} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_g^{\mathrm{UBM}}, \boldsymbol{\Sigma}_g^{\mathrm{UBM}})}. \tag{2}$$

These probabilities are used to obtain the sufficient statistics for the mean vectors given by

$$n_i = \sum_{t=1}^{T} P(i \mid \mathbf{x}_t), \tag{3}$$

$$\mathbf{E}_i = \frac{1}{n_i} \sum_{t=1}^{T} P(i \mid \mathbf{x}_t) \mathbf{x}_t. \tag{4}$$

In the second stage, sufficient statistics obtained from the input acoustic feature vector sequence by the maximum likelihood criterion are used to update the UBM mean vector using the following equation.

$$\boldsymbol{\mu}_i^{\mathrm{MAP}} = \omega_i \mathbf{E}_i + (1 - \omega_i) \boldsymbol{\mu}_i^{\mathrm{UBM}}, \tag{5}$$

where $\boldsymbol{\mu}_i^{\mathrm{UBM}}$ and $\boldsymbol{\mu}_i^{\mathrm{MAP}}$ denote the i -th UBM mean vector and the i -th updated mean vector, respectively. The variable, ω_i, is used to determine the balance between the sufficient statistics and UBM mean vector. ω_i is given by $\omega_i = n_i / (n_i + r)$, where r, called the relevance factor, was set to 16 in this study.

2.2 Support Vector Machines

SVM [9] is a maximum margin classifier, which is a state-of-the-art discriminative classification algorithm used in the speaker verification area. Basically, the SVM is a two-class classifier with a linear hyperplane decision boundary. The two classes consist of positive and negative classes, which are labeled +1 and -1, respectively. In particular, in the speaker verification task, positive and negative classes correspond to a registered speaker and impostor, respectively.

The SVM discriminant function is given as

$$f(\mathbf{x}) = \sum_{l=1}^{L} \alpha_l t_l K(\mathbf{x}, \mathbf{x}_l) + d, \tag{6}$$

where L is the number of support vectors, $\mathbf{x}_l$ is the l -th support vector, and $t_l \in \{1, -1\}$ is the corresponding class label. α_l and d are the SVM parameters obtained from a training set by an optimization process. $\mathbf{x}$ is an input GMM supervector. The kernel function $K(\cdot, \cdot)$ must satisfy the Mercer condition so $K(\cdot, \cdot)$ can be expressed as an inner product form given as

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^{\mathrm{T}} \phi(\mathbf{y}), \tag{7}$$

where $\phi(\mathbf{x})$ is a mapping function from the input space to the SVM feature space. Fig. 1 shows the overall process of the MAP adaptation-based GMM-SVM speaker verification system.

3. Sparse Maximum A Posteriori Adaptation

The MAP adaptation applied in the GMM-SVM

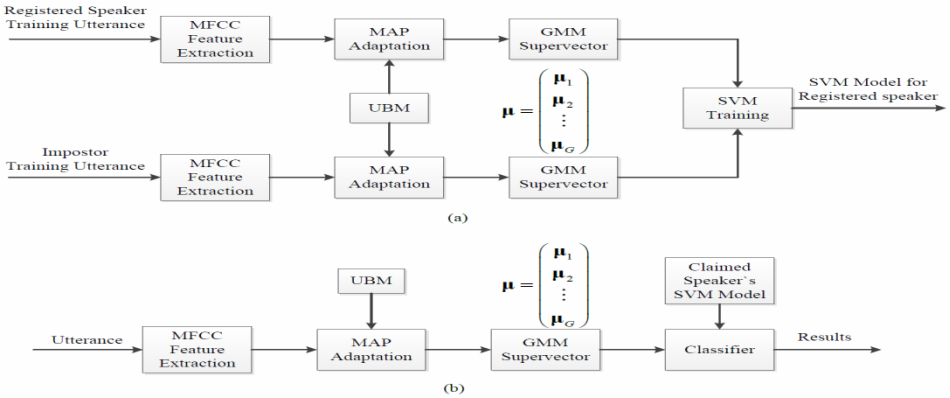

Fig. 1. Overall process of the MAP adaptation based GMM-SVM speaker verification system: (a) training phase, (b) test phase.

speaker verification system utilizes the conjugate Bayesian priors for the Gaussian parameters (normal distribution for the mean vectors) to adapt the parameters of the UBM with a given utterance. The number of parameters to be estimated would be very large compared to the length of the given utterance, and although the Bayesian prior provides certain smoothing, small variability still remains, which works as adaptation noise. Sparse MAP adaptation [8, 9] was proposed to remove the adaptation noise by introducing sparseness constraint to the MAP criteria.

Assuming that the covariance matrices in the GMM are diagonal and the adapting parameters are limited to the mean vectors, the sparse constrained MAP problem, where only N parameters are allowed to update, can be expressed as

$$\begin{aligned} &\max_{\boldsymbol{\Theta}} \sum_{i=1}^{G}\sum_{d=1}^{D}(n_i+r)L(\mathbf{X};\mu_{id}\in\{\mu_{id}^{\mathrm{MAP}},\mu_{id}^{\mathrm{UBM}}\}) \\ &\text{subject to } N=\sum_{i=1}^{G}\sum_{d=1}^{D}\left\|\mu_{id}-\mu_{id}^{\mathrm{UBM}}\right\|_0, \end{aligned} \tag{8}$$

where $\mathbf{X}=\{\mathbf{x}_1,\cdots,\mathbf{x}_T\}$ denotes the acoustic feature vector sequence, G is the number of mixtures, D is the dimension of the acoustic feature vector, $\boldsymbol{\Theta}=\{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_G\}$ is a set of adapting parameters, $\left\|\mu_{id}-\mu_{id}^{\mathrm{UBM}}\right\|_0\in\{0,1\}$ is the L0 norm, and i and d denote the index of the mixture and the index of the acoustic feature vector dimension, respectively. $L(\cdot)$ is a log likelihood function that calculates the likelihood of feature vectors with the given model parameters. The problem can be solved exactly because the objective function and constraints in Eq. (8) are composed of direct sums over i and d. The optimal solution for N can be obtained by a greedy search for the parameters that increase the objective function the most. For example, if N is equal to 1, it is clear that the single parameter change that causes the largest increase in the objective function is optimal.

Alternatively, a maximization process can be considered using the Lagrangian function given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta};\tau)=&\sum_{i=1}^{G}\sum_{d=1}^{D}(n_i+r)L(\mathbf{X};\mu_{id}\in\{\mu_{id}^{\mathrm{MAP}},\mu_{id}^{\mathrm{UBM}}\}) \\ &-\tau\Big(\sum_{i=1}^{G}\sum_{d=1}^{D}\left\|\mu_{id}-\mu_{id}^{\mathrm{UBM}}\right\|_0-N\Big), \end{aligned} \tag{9}$$

where τ is the Lagrangian coefficient. Eq. (9) could not give the global optimum because the objective function is not convex. On the other hand, for a fixed τ, Eq. (9) fully decouples across i, d and each sub-problem given by

$$\max_{\mu_{id}} \left\{(n_i+r)L(\mathbf{X};\mu_{id}\in\{\mu_{id}^{\mathrm{MAP}},\mu_{id}^{\mathrm{UBM}}\})-\tau\left\|\mu_{id}-\mu_{id}^{\mathrm{UBM}}\right\|_0\right\}, \tag{10}$$

can be solved independently. In addition, $\mu_{id}\in\{\mu_{id}^{\mathrm{MAP}},\mu_{id}^{\mathrm{UBM}}\}$ has only two options and each sub-problem can be solved directly. Because $L(\cdot)$ of Eq. (10) denotes the log likelihood, the L0 norm in Eq. (10) acts as a sparsity promoting regularizer. For computational benefit, it is useful to minimize the following equation instead of Eq. (10).

$$\begin{aligned} F(\mathbf{X};\mu_{id}\in\{\mu_{id}^{\mathrm{MAP}},\mu_{id}^{\mathrm{UBM}}\})=& \\ -2L(\mathbf{X};\mu_{id}\in\{\mu_{id}^{\mathrm{MAP}},\mu_{id}^{\mathrm{UBM}}\})&+\alpha\left\|\mu_{id}-\mu_{id}^{\mathrm{UBM}}\right\|_0 \end{aligned} \tag{11}$$

$$\alpha=\frac{2\tau}{n_i+r} \tag{12}$$

The optimization solution for Eq. (11) is given by Algorithm 1. According Algorithm 1, the sparse MAP adaptation method can be implemented with few additional operations compared to the MAP adaptation.

4. Experiments

Experiments were performed on the TIMIT database

Algorithm 1 Solution to sparse MAP adaptation

Input: μ_{id}^{MAP}, μ_{id}^{UBM}, $(\sigma^2)_{id}^{\text{UBM}}$, τ, n_i, and r

Output: μ_{id}

Compute $\alpha = 2\tau/(n_i + r)$

Compute Lagrangian objectives:

$$F_1 \leftarrow \frac{(\mu_{id}^{\text{UBM}} - \mu_{id}^{\text{MAP}})}{(\sigma^2)_{id}^{\text{UBM}}} + 1 + \log(2\pi(\sigma^2)_{id}^{\text{UBM}})$$

$$F_2 \leftarrow 1 + \log(2\pi(\sigma^2)_{id}^{\text{UBM}}) + \alpha$$

if $F_1 < F_2$ then $\mu_{id} = \mu_{id}^{\text{UBM}}$

else $\mu_{id} = \mu_{id}^{\text{MAP}}$

[11], which had been collected at a 16 kHz sampling rate and 16-bit resolution. TIMIT contains a total of 6300 sentences, which are composed of 10 sentences spoken by each of 630 speakers and each sentence is composed of approximately 2 to 3 seconds long speech data. 1280 sentences of 160 speakers (80 male and 80 female speakers) were used to construct the UBM without the SA1 and SA2 utterances, and 100 randomly selected registered speakers and 330 impostors were used. The others in the database were composed of background speakers. For the registered speakers, 8 utterances per speaker were used for SVM training and the remaining 2 utterances were used for the test. A 5-fold cross-validation was performed because the amount of the target trial was too small. This setup resulted in 1000 target trials and 3300 non-target trials. The performance evaluation was measured by the equal error rate (EER) and detection error tradeoff (DET) curves obtained by varying the decision threshold.

The acoustic feature extraction were performed with 13-dimensional Mel-frequency cepstral coefficients (MFCCs) with a 0.97 pre-emphasis coefficient, 10 ms frame shift size, and 20 ms Hamming window. The first derivatives of the MFCCs were then added to the acoustic feature to capture the dynamic characteristics of speech. An energy-based voice activity detector was followed to remove the silences unrelated to the speaker dependent characteristics.

Table 1 lists the EER results that were obtained in terms of the number of mixtures and the Lagrangian coefficient τ. In the sparse MAP-SVM system, as τ decreases, the speaker-dependent GMM supervectors are more sparsely updated from the UBM. In Table 1, the sparse MAP-SVM system always showed better performance than the conventional MAP-SVM system with the proper settings for τ and the best performance occurred in the sparse MAP-SVM system with 512 mixtures and $\tau = 0.5$. The maximum relative EER reduction was 42.6% at the setting. Note that the verification results are degraded with 1024 mixtures because the data used for adaptation is not sufficient to adapt all the mean vectors. Furthermore, the range for τ showing a higher EER than MAP-SVM in

Table 1 become wider with increasing number of mixtures, which suggests that adaptation noise also increases. Fig. 2 shows the DET curves of the MAP-SVM

Table 1. Equal error rate evaluation results with various mixtures of 128, 256, 512, and 1024.

Number of mixtures		128	256	512	1024
MAP-SVM		**2.50%**	**2.61%**	**2.58%**	**2.70%**
Sparse MAP-SVM	τ=0.001	27.36%	18.90%	11.39%	8.90%
	0.01	21.50%	11.70%	6.90%	5.80%
	0.1	6.90%	3.30%	2.18%	**1.79%**
	0.5	2.90%	**1.80%**	**1.48%**	**1.79%**
	1	2.60%	1.90%	1.50%	2.00%
	2	2.09%	1.90%	1.70%	2.39%
	5	**2.00%**	2.20%	2.09%	2.39%
	10	2.20%	2.50%	2.45%	2.50%
	100	2.50%	2.61%	2.58%	2.70%

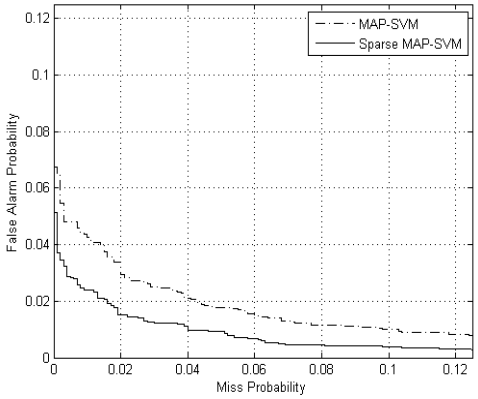

Fig. 2. DET curves of the MAP-SVM and sparse MAP-SVM with 1024 mixtures.

system and the sparse MAP-SVM system with $\tau = 0.5$ for 1024 mixtures. As shown in the figure, the sparse MAP-SVM system also shows better performance than the conventional MAP-SVM.

5. Conclusion

For some applications with a limited length of input utterance compared to the parameters to be adapted, the MAP adaptation method can cause adaptation noise. This paper introduced the sparse MAP adaptation to the GMM-SVM-based text-independent speaker verification system to mitigate the adaptation noise. From pseudo-code, it was shown that the sparse MAP adaptation requires few additional operations compared to the MAP adaptation. The performance of the sparse MAP-SVM system was improved compared to the conventional MAP-SVM system, with a maximum 42.6% relative EER reduction on the TIMIT database. Nevertheless, future studies will be needed to find an automatic method for selecting τ and apply the proposed method to state-of-the-art techniques as a preprocessor generating speaker-dependent GMM supervectors instead of the MAP adaptation.

Acknowledgement

This work was supported by the IT R&D program of MSIP/KEIT. [10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning]

References

[1] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, pp. 308-311, May 2006. Article (CrossRef Link)

[2] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, 2006, pp. 97–100. Article (CrossRef Link)

[3] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Pittsburgh, PA, Sep. 2006. Article (CrossRef Link)

[4] R. Vogt, S. Kajarekar, and S. Sridharan, "Discriminat NAP for SVM speaker recognition," in *Proc. IEEE Odyssey: Speaker Lang. Recogn. Workshop*, Stellenbosch, South Africa, Jan. 2008. Article (CrossRef Link)

[5] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Improved SVM speaker verification through data-driven background dataset collection," in *Proc. IEEE ICASSP*, Taipei, Taiwan, 2009, pp. 4041–4044. Article (CrossRef Link)

[6] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Dig. Signal Process.*, vol. 10, pp. 19-41, 2000. Article (CrossRef Link)

[7] J. Gauvain and C.-H. Lee, "Maximum *a posterior* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. *2*, pp. 291-298, Apr. 1994 Article (CrossRef Link)

[8] P. A. Olsen, J. Huang, V. Goel, S. J. Rennie, "Sparse maximum a posteriori adaptation," *in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 53-58, 2011. Article (CrossRef Link)

[9] P. A. Olsen, J. Huang, S. J. Rennie, and V. Goel, "Affine invariant sparse maximum a posteriori adaptation," in Proc. ICASSP, 2012, pp. 4317 – 4320. Article (CrossRef Link)

[10] B. Scholkopf and A. J. Smola, *Learning with kernels*, The MIT Press, 2002. Article (CrossRef Link)

[11] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Darpa TIMIT: acoustic-phonetic continuous speech corpus CD-ROM," *LDC catalog number LDC93s1*, 1993. Article (CrossRef Link)

Younggwan Kim received the B.S. degree from the Department of Radio Frequency Engineering, Hanbat National University, Daejeon, Korea, in 2008 and the M.S. degree from the Department of Information and Communications Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2010. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, KAIST. His research interests include signal processing for speech and speaker recognition, content-based audio indexing and classification, and emotional speech classification.

Jaeyoung Roh received the B.S. degree from the School of Electronics and Electrical Engineering, Sungkyunkwan University, Suwon, Korea, in 2011 and the M.S. degree from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2013. His research interests include signal processing for speech and speaker recognition.

Hoirin Kim received the B.S. degree from the Department of Electronics Engineering, Hanyang University, Seoul, Korea in 1984, and the M.S. and Ph.D. degrees from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1987 and 1992, respectively. From October 1987 to December 1999, he was a Senior Researcher in the Spoken Language Processing Lab, Electronics and Telecommunications Research Institute (ETRI), Korea. From June 1994 to May 1995, he was on leave to the ATR-ITL, Kyoto, Japan, as a Visiting Researcher. From January 2000 to February 2009, he was an Associative Professor at the Information and Communications University (ICU), Korea. From July 2006 to June 2007, he was on leave to the INC, University of California, San Diego, as a Visiting Scholar. Since March 2009, he has been a Professor in the Department of Electrical Engineering, KAIST. His research interests include signal processing for speech and speaker recognition, audio indexing and retrieval, and spoken language processing.

Copyrights © 2013 The Institute of Electronics Engineers of Korea