한국음향학회지 제32권 제2호 pp. 174～180 (2013)
The Journal of the Acoustical Society of Korea Vol.32, No.2 (2013)
http://dx.doi.org/10.7776/ASK.2013.32.2.174

pISSN : 1225-4428
eISSN : 2287-3775

Decision Tree Based Context Clustering with Cross Likelihood Ratio for HMM-based TTS

HMM 기반의 TTS를 위한 상호유사도 비율을 이용한 결정트리 기반의 문맥 군집화

Chi-Sang Jung† and Hong-Goo Kang
(정치상†, 강홍구)

School of Electrical and Electronic Engineering, Yonsei University

(Received December 3, 2012; revised December 28, 2012; accepted January 27, 2013)

ABSTRACT: This paper proposes a decision tree based context clustering algorithm for HMM-based speech synthesis systems using the cross likelihood ratio with a hierarchical prior (CLRHP). Conventional algorithms tie the context-dependent HMM states that have similar statistical characteristics, but they do not consider the statistical similarity of split child nodes, which does not guarantee the statistical difference between the final leaf nodes. The proposed CLRHP algorithm improves the reliability of model parameters by taking a criterion of minimizing the statistical similarity of split child nodes. Experimental results verify the superiority of the proposed approach to conventional ones.

Key words: HMM-based TTS, Decision tree based context clustering, cross likelihood ratio with hierarchical prior (CLRHP)
PACS numbers: 43.72.Ja

초 록: 본 논문은 HMM 기반의 TTS 시스템을 위하여 상호유사도 비율을 이용한 결정트리 기반의 문맥 군집화 알고리즘을 제안한다. 기존의 알고리즘들은 유사한 통계적 특성을 가지는 문맥종속 HMM을 하나로 묶고 있다. 그러나 기존의 알고리즘들은 결정트리의 나누어진 노드간의 통계적 유사도를 고려하지 않음으로 인하여 최종 노드 사이의 통계적인 차이를 보장하지 못한다. 제안한 알고리즘은 분리된 노드들 간의 통계적 유사도를 최소화하여 모델 파라미터의 신뢰도를 향상시킨다. 실험 결과를 통해 제안한 알고리즘이 기존의 알고리즘들에 비해 우수한 성능을 나타낸다는 것을 확인할 수 있다.

핵심용어: HMM 기반 음성합성기, 결정트리 기반 문맥 군집 알고리즘, 상호 유사 비율

I. Introduction

The training process of HMM-based speech synthesis system requires a context-dependent approach.[1] As the various combination of context information is utilized, the size of context-dependent HMMs increases unlimitedly. Since the amount of speech database is insufficient to represent all combination of context information, however, it is very difficult to obtain model parameters reliably. To relieve the problem by merging similar training sets, various types of context clustering techniques have been proposed.[2-6]

The Minimum Description Length (MDL)-based context clustering algorithm is the most popular one.[2-3] The HMM parameters, i.e. mean vector and covariance matrix, having the similar statistics are tied to the same context-dependent HMM states by the contextual dependent questions. However, since the MDL algorithm uses the Maximum Likelihood (ML) criterion, it has an overfitting problem if the amount of training data at any node is insufficient and outliers occur at any node.

†**Corresponding author:** Chi-Sang Jung (jtoctos@dsp.yonsei.ac.kr)
School of Electrical and Electronic Engineering, Yonsei University, Shinchon-dong, Seodaemun-gu, Seoul 120-749, Republic of Korea
(Tel: 82-2-2123-4534; Fax: 82-2-364-4870)

In order to solve the overfitting problem, the cross validation for node splitting and stopping criterion was proposed.[4,5] Since it is still based on the ML criterion, however, the estimated model parameters are still not reliable if the amount of training data at any node is insufficient. A hierarchical prior-based context clustering algorithm is proposed to overcome the insufficiency of training data.[6] The hierarchical prior that is similar to Structural Maximum A Posteriori (SMAP) estimation[7] regularizes a parameter estimation process, thus it is possible to estimate reliable model parameters.

In the context clustering algorithms, the characteristics of model parameters at each node vary depending on the best chosen contextual question. Moreover, the effect caused by selecting any of the best question at the root node is propagated to the hierarchical tree structure afterwards. Therefore, it is important for designing a splitting criterion in the training process.

In the conventional algorithms, the node splitting criterion only depends on the likelihood ratio between the current node and child nodes to be split. In other words, a splitting criterion tries to maximize the summation of log-likelihood to each split child node. However, since the criterion does not consider the level of similarity between each other child node, it does not guarantee an important point such that the characteristics of model parameters in each other child node need to be statistically different.

This paper proposes a novel splitting criterion including the Cross Likelihood Ratio between split child nodes with their Hierarchical Prior (CLRHP). To minimize the similarity between each other child node, the proposed CLRHP algorithm includes the cross likelihood ratio into the splitting criterion. Furthermore, the cross likelihood ratio is normalized by the log-likelihood of each child node using the hierarchical prior, which results in the regularization process to the statistics of parent node. Subjective and objective test results show that the performance of the proposed context clustering algorithm is superior to the conventional ones.

II. Conventional Decision Tree Based Context Clustering Algorithms

2.1 MDL-based Context Clustering

The MDL-based context clustering keeps balancing between limited amount of speech database and unlimited combination of context information.[3] The approach takes a top-down clustering approach such that it maximizes the likelihood of model parameters to the training data. In the HMM-based TTS system, an HMM state-level clustering is adopted.

When the clustered node S is divided into S_{q+} and S_{q-} by a question q, the MDL-based criterion is defined as:

$$\Delta_q = \left\{ L(S_{q+}) + L(S_{q-}) \right\} - L(S) - \alpha \frac{N}{2} \log \Gamma(S), \quad (1)$$

where denotes a weight of penalty term, and N is the number of parameters increased by the split. The total state occupancy count at the node S, $\Gamma(S)$, is defined as:

$$\Gamma(S) = \sum_{t=1}^{T} \sum_{m \in \mathbf{M}_S} \gamma_m(t), \quad (2)$$

where T denotes the number of frames in the training data, M_S denotes a set of HMM states clustered to the node S. $\gamma_m(t)$ denotes the posteriori probability of an HMM state m for an observation at frame t, $\mathbf{o}_t$. The log-likelihood at node S to the associated training data is defined as:

$$\begin{aligned} L(S) &= \sum_{t=1}^{T} \sum_{m \in \mathbf{M}_S} \gamma_m(t) \log(\mathbf{o}_t; \boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S) \\ &= -\frac{1}{2} \sum_{t=1}^{T} \sum_{m \in \mathbf{M}_S} \gamma_m(t) \left\{ K + K \log(2\pi) + \log|\boldsymbol{\Sigma}_S| \right\}, \end{aligned} \quad (3)$$

where K denotes the dimensionality of observation vector. $\boldsymbol{\mu}_S$ and $\boldsymbol{\Sigma}_S$ denote the mean and covariance of the leaf node S, respectively.

To determine whether a current node is split or not, the optimum question $\hat{q}$ that maximizes Δq is chosen at first. Then, if $\Delta\hat{q} < 0$, the current node is not split, otherwise the node S is divided into child nodes, S_{q+} and S_{q-}. This process is repeatedly carried out until there remains no nodes to be split.

Note that the model parameters trained by the ML estimator are sensitive to the outliers if the amount of training data is insufficient. Furthermore, controlling the penalty term is not easy to determine an appropriate tree size. To overcome the problems of MDL-based context clustering, the Cross Validation (CV) and the Cross Validation based on the hierarchical prior similar to Structural Maximum a Posteriori estimation (CVSMAP) based context clustering algorithms are proposed in.[5-6]

2.2 Decision Tree Based Context Clustering Based on Cross Validation and Hierarchical Prior

In cross validation, the training data D is divided into K-folds $D^{(1)},...,D^{(K)}$, where $D = \bigcup_{i=1}^{K} D^{(i)}$ and $D^{(i)} \cap D^{(j)} = 0$ for any i and j. The occupancy counts and the first and second order statistics of $D^{(k)}$ associated with the node S are written as:

$$\begin{aligned}
\Gamma_S^{(k)} &= \sum_{t\in D^{(k)}} \sum_{m\in M_S} \gamma_m(t), \\
\mathbf{v}_S^{(k)} &= \sum_{t\in D^{(k)}} \sum_{m\in M_S} \gamma_m(t)\mathbf{o}_t, \\
\mathbf{\Omega}_S^{(k)} &= \sum_{t\in D^{(k)}} \sum_{m\in M_S} \gamma_m(t)\mathbf{o}_t\mathbf{o}_t^{\cdot},
\end{aligned} \tag{4}$$

where $\Gamma_S^{(k)}, \mathbf{v}_S^{(k)}$ and $\mathbf{\Omega}_S^{(k)}$ denote the occupancy counts, the first, and second statistics, respectively. For the k-th fold, the subsets $D^{(\bar{k})} = \bigcup_{i\neq k} D^{(i)}$ are used for the ML estimator, where $\bar{k} = \{1,2,...,k-1,k+1,...,K\}$. The mean vector $\boldsymbol{\mu}_S^{(k)}$ and covariance matrix $\mathbf{\Sigma}_S^{(k)}$ at the node S are estimated as:

$$\boldsymbol{\mu}_S^{(k)} = \frac{\mathbf{v}_S^{(\bar{k})}}{\Gamma_S^{(\bar{k})}}, \qquad \mathbf{\Sigma}_S^{(k)} = \frac{\mathbf{\Omega}_S^{(\bar{k})}}{\Gamma_S^{(\bar{k})}} - \boldsymbol{\mu}_S^{(k)}\boldsymbol{\mu}_S^{(k)^{\cdot}}, \tag{5}$$

where

$$\begin{aligned}
&\Gamma_S^{(\bar{k})} = \sum_{i\neq k} \Gamma_S^{(i)} + \tau_S, \\
&\mathbf{v}_S^{(\bar{k})} = \sum_{i\neq k} \mathbf{v}_S^{(i)} + \tau_S \frac{\mathbf{v}_{S^P}^{(\bar{k})}}{\Gamma_{S^P}^{(\bar{k})}}, \ \ \mathbf{\Omega}_S^{(\bar{k})} = \sum_{i\neq k} \mathbf{\Omega}_S^{(i)} + \tau_S \frac{\mathbf{\Omega}_{S^P}^{(\bar{k})}}{\Gamma_{S^P}^{(\bar{k})}}.
\end{aligned} \tag{6}$$

S^P denotes the parent node of S and τ_S denotes the regularization factor for the prior statistics of S. Note that S equals zero in the CV-based algorithm. The use of hierarchical priors regularizes parameter estimation, thus reduces the overfitting problem.

From the first and second statistics, and equation (3), the log-likelihood of the model at the node S given the evaluation subset $D^{(k)}$ can be calculated as

$$\begin{aligned}
&L(\lambda_S^{(k)}; D_S^{(k)}) \\
&= -\frac{1}{2}\Big\{\Gamma_S^{(k)} \log\left(2\pi\left|\mathbf{\Sigma}_S^{(k)}\right|\right) + \mathrm{tr}\left(\mathbf{\Omega}_S^{(k)}\mathbf{\Sigma}_S^{(k)^{-1}}\right) \\
&\quad - 2\boldsymbol{\mu}_S^{(k)^{\text{빈}}}\mathbf{\Sigma}_S^{(k)^{-1}}\mathbf{v}_S^{(k)} + \Gamma_S^{(k)}\boldsymbol{\mu}_S^{(k)}\ \mathbf{\Sigma}_S^{(k)^{-1}}\boldsymbol{\mu}_S^{(k)}\Big\},
\end{aligned} \tag{7}$$

where $\lambda_S^{(k)} = \{\boldsymbol{\mu}_S^{(k)}, \mathbf{\Sigma}_S^{(k)}\}$. This process is repeated over K-folds. Then, the CV or CVSMAP log-likelihood at the node S is calculated by summing the likelihood of each subset :

$$L(S;D) = \sum_{k=1}^{K} L(\lambda_S^{(k)}; D_S^{(k)}). \tag{8}$$

In the CVSMAP-based algorithm, the prior statistics for the root node S_0 are set to

$$\Gamma_{S_0^P}^{(\bar{k})} = 1, \ \ \mathbf{v}_{S_0^P}^{(\bar{k})} = \mathbf{0}, \ \ \mathbf{\Omega}_{S_0^P}^{(\bar{k})} = \mathbf{I}, \ \ k = 1,...,K. \tag{9}$$

The splitting and stopping criterion in the CV or CVSMAP based algorithm are same as the MDL method except the penalty term, while the log-likelihood is calculated by equation (7) and (8).

III. Proposed Context Clustering Algorithm

In the context clustering algorithm, determining the splitting criterion is important to make the model parameters be statistically inequivalent. In other words, the clustering algorithm is efficient if each other clustered models have statistically different characteristics. Typically, the splitting criteria try to maximize the summation of log-likelihoods of split child nodes. However, the conventional splitting criteria do not guarantee the fact that the model parameters of each other child nodes are statistically different because the criteria do not include the statistical similarity measure between each other child nodes. In this section, a novel splitting criterion with the CLRHP is proposed, which includes the similarity between each other child nodes into the criterion. Therefore, the CLRHP criterion not only maximizes the likelihood at the child nodes, but also minimizes the similarity of model parameters using the cross likelihood ratio.

3.1 Normalized Cross Likelihood Ratio

To describe the concept of cross likelihood ratio, a Normalized Cross Likelihood Ratio (NCLR)[8] should be defined first. The NCLR means a distance measure between two models having Gaussian distribution. Given two Gaussian models, M_i and M_j , the NCLR distance is defined as :

$$\begin{aligned} &NCLR\left(M_i, M_j\right) \\ &= \frac{1}{N_i}\log\left(\frac{\mathbb{N}(\mathbf{o}_i; M_i)}{\mathbb{N}(\mathbf{o}_i; M_j)}\right) + \frac{1}{N_j}\log\left(\frac{\mathbb{N}(\mathbf{o}_j; M_j)}{\mathbb{N}(\mathbf{o}_j; M_i)}\right), \end{aligned} \tag{10}$$

where N_i and N_j are the number of data $\mathbf{O}_i$ and $\mathbf{O}_j$, respectively. $\mathbb{N}(\mathbf{o}_i; M_j)$ and $\mathbb{N}(\mathbf{o}_j; M_i)$ denote the cross likelihood of two Gaussian models M_i and M_j, and they are normalized by their own likelihood $\mathbb{N}(\mathbf{o}_i; M_i)$and $\mathbb{N}(\mathbf{o}_j; M_j)$, respectively.

3.2 Node Splitting Criterion with Cross Likelihood Ratio Considering the Hierarchical Prior

From the log-likelihood ratio used for the cross validation with the hierarchical prior, and the modified normalized cross likelihood ratio, we propose a node splitting criterion using the CLRHP. The CLRHP-based splitting and stopping criterion is defined as:

$$\begin{aligned} \Delta_q = &\left\{L(S_{q+}; D) + L(S_{q-}; D)\right\} - L(S; D) \\ &+ CLR_{HP}\left(S; D\right), \end{aligned} \tag{11}$$

where q is the best question at each node. $L(S_{q+}; D)$, $L(S_{q-}; D)$, and $L(S; D)$ are calculated by the equation given in (7) and the node statistics based on the hierarchical priors, i.e. equation (5), and (6). The CLRHP term $CLR_{HP}(S; D)$ is defined in the equation (12). The way of computing the log-likelihoods is given in equation (13), which means that $CLR_{HP}\left(S; D\right)$ is calculated by the occupancy count and the first and second statistics at each

$$CLR_{HP}\left(S;D\right) = \sum_{k=1}^{K}\left[\frac{1}{\Gamma_{S_{q+}}^{(k)}}\left\{L\left(S_{q+}^{(k)}, D_{q+}^{(k)}; S\right) - L\left(S_{q+}^{(k)}, D_{q+}^{(k)}; S_{q-}\right)\right\} + \frac{1}{\Gamma_{S_{q-}}^{(k)}}\left\{L\left(S_{q-}^{(k)}, D_{q-}^{(k)}; S\right) - L\left(S_{q-}^{(k)}, D_{q-}^{(k)}; S_{q+}\right)\right\}\right]. \tag{12}$$

$$\begin{aligned} L\left(S_{q+}^{(k)}, D_{q+}^{(k)}; S_{q-}\right) &= -\frac{1}{2}\left[Tr\left(\left(\mathbf{\Omega}_{S_{q+}}^{(k)} - \boldsymbol{\mu}_{S_{q-}}\mathbf{v}_{S_{q+}}^{(k)T} - \mathbf{v}_{S_{q+}}^{(k)}\boldsymbol{\mu}_{S_{q-}}{}^{T} + \boldsymbol{\mu}_{S_{q-}}\boldsymbol{\mu}_{S_{q-}}{}^{T}\right)\mathbf{\Sigma}_{S_{q-}}^{-1}\right) + \Gamma_{S_{q+}}^{(k)}\left(K\log(2\pi) + \log\left|\mathbf{\Sigma}_{S_{q-}}\right|\right)\right], \\ L\left(S_{q-}^{(k)}, D_{q-}^{(k)}; S_{q+}\right) &= -\frac{1}{2}\left[Tr\left(\left(\mathbf{\Omega}_{S_{q-}}^{(k)} - \boldsymbol{\mu}_{S_{q+}}\mathbf{v}_{S_{q-}}^{(k)T} - \mathbf{v}_{S_{q-}}^{(k)}\boldsymbol{\mu}_{S_{q+}}{}^{T} + \boldsymbol{\mu}_{S_{q+}}\boldsymbol{\mu}_{S_{q+}}{}^{T}\right)\mathbf{\Sigma}_{S_{q+}}^{-1}\right) + \Gamma_{S_{q-}}^{(k)}\left(K\log(2\pi) + \log\left|\mathbf{\Sigma}_{S_{q+}}\right|\right)\right], \\ L\left(S_{q+}^{(k)}, D_{q+}^{(k)}; S\right) &= -\frac{1}{2}\left[Tr\left(\left(\mathbf{\Omega}_{S_{q+}}^{(k)} - \boldsymbol{\mu}_{S}\mathbf{v}_{S_{q+}}^{(k)T} - \mathbf{v}_{S_{q+}}^{(k)}\boldsymbol{\mu}_{S}{}^{T} + \boldsymbol{\mu}_{S}\boldsymbol{\mu}_{S_q}{}^{T}\right)\mathbf{\Sigma}_{S}^{-1}\right) + \Gamma_{S_{q+}}^{(k)}\left(K\log(2\pi) + \log\left|\mathbf{\Sigma}_{S}\right|\right)\right], \\ L\left(S_{q-}^{(k)}, D_{q-}^{(k)}; S\right) &= -\frac{1}{2}\left[Tr\left(\left(\mathbf{\Omega}_{S_{q-}}^{(k)} - \boldsymbol{\mu}_{S}\mathbf{v}_{S_{q-}}^{(k)T} - \mathbf{v}_{S_{q-}}^{(k)}\boldsymbol{\mu}_{S}{}^{T} + \boldsymbol{\mu}_{S}\boldsymbol{\mu}_{S}{}^{T}\right)\mathbf{\Sigma}_{S}^{-1}\right) + \Gamma_{S_{q-}}^{(k)}\left(K\log(2\pi) + \log\left|\mathbf{\Sigma}_{S}\right|\right)\right]. \end{aligned} \tag{13}$$

node. In summary, the proposed CLRHP criterion utilizes the occupancy count and the first and second statistic considering the cross validation and hierarchical priors.

Ⅳ. Performance Evaluation

4.1 Experimental Setup

The performance of the proposed algorithm is compared to those of conventional context clustering algorithms. At first, a Korean HMM-based TTS system is constructed.[9-11] Around three thousand Korean utterances are recorded by a professional male speaker for training. Fifty sentences which are not included in the training set are used for the test. The sampling frequency is set to 16 kHz, and there are 181,734 combinations of context information. A grapheme-to- phoneme (G2P) converter is implemented by following the Korean standard pronunciation grammar and the context information labeling program. Sixteenth-order LSFs are used for the spectral parameter and twenty third-order excitation parameters including F0 are used for excitation parameters.[11]

For the MDL-based context clustering algorithm, a penalty factor is tuned to 0.2~2.0 with an interval of 0.2. For the CV-based context clustering algorithm, the number of folds, K, is set to 5. In the CVSMAP-based context clustering algorithm, the regularization parameter for the CVSMAP is replaced to considering the ratio of occupancy counts between the current and parent nodes. K and $\alpha_s^{(\bar{k})}$ are also applied to the proposed CLRHP-based context clustering algorithm.

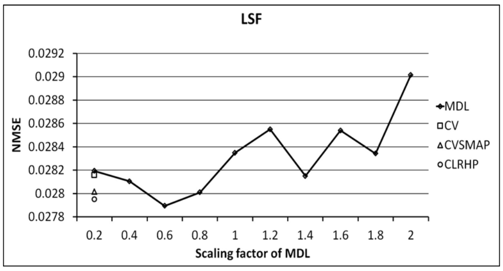

Fig. 1. NMSE of LSF for each clustering algorithm.

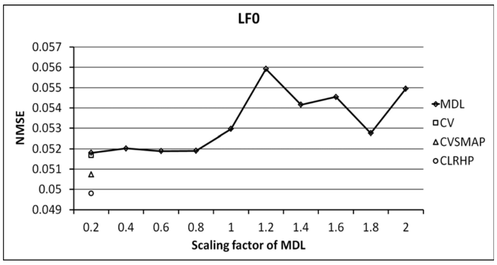

Fig. 2. NMSE of LF0 for each clustering algorithm.

$$\alpha_s^{(\bar{k})} = \frac{\sum_{i \neq k} \Gamma_S^{(i)}}{\Gamma_{S^P}} \quad . \tag{14}$$

4.2 Analysis on the Trainability of Context Clustering Algorithms

In order to analyze the trainability of each context clustering algorithm, we investigate the trainability of the spectral and excitation features. The NMSE is defined as a normalized error between excitation parameters extracted from the original speech and those generated from the trained HMMs .

$$NMSE = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{\sum_{n=1}^{N}\left(x_{org}(t,n) - x_{syn}(t,n)\right)^2}{\sum_{n=1}^{N}\left(x_{org}(t,n)\right)^2}} \quad . \tag{15}$$

Fig. 1 and 2 show the average NMSE of LSF and LF0, respectively. In Fig. 1, the NMSE value of MDL is varying depending on the scaling factor of penalty term. The NMSE values of CV, CVSMAP, and CLRHP are similar to the minimum value of NMSE of MDL. It means that the decision tree of LSF does not have a large variation depending on the type of clustering algorithm. The error reduction of spectral parameter depending on the clustering algorithm is less than one of the excitation parameter.

In Fig. 2, the error of excitation parameters, LF0, is reduced by the proposed CLRHP algorithm compared to the conventional algorithms. Note that the value of reduced NMSE is not large because we utilize the large amount of speech database about 4 hours for training the context-dependent HMMS. Nevertheless, the CLRHP has

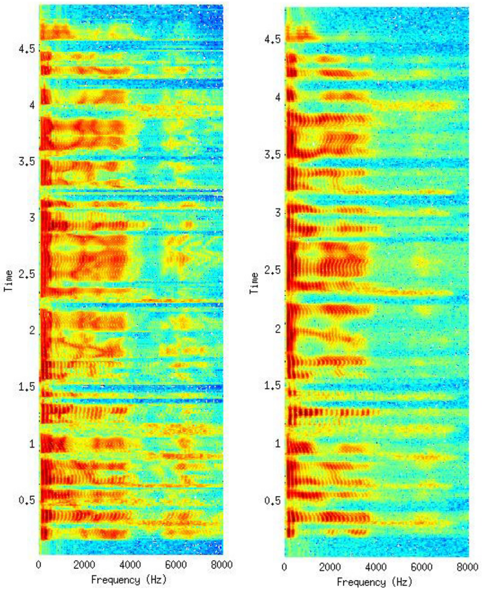

Fig. 3. Original (left) and proposed CLRHP (right) speech spectra.

the smallest NMSE among all clustering algorithms. Since the excitation and spectral parameters are independently trained and clustered by each other decision tree, the results of NMSE shows that the clustering algorithm has more effect on the excitation parameters than the spectral parameters. In Fig. 3, the spectrogram synthesized by the proposed CLRHP is compard to one of the original speech.

4.3 Objective Test Results

In order to evaluate the objective quality of synthesized speech by each context clustering algorithm, a log spectral distance (LSD) between the original and generated speech is measured in the speech duration. Fig. 4 represents the LSD values for each context clustering algorithm. It is clear that the spectral distortion of speech synthesized by the proposed CLRHP algorithm has lower value than the conventional context clustering algorithms, i.e. MDL, CV, and CVSMAP.

4.4 Subjective Test Results

In order to measure the subjective speech quality, the A/B/X preference test is also conducted. Fifteen experts in speech signal processing field provide their preference after listening randomized utterances synthesized by four methods. In this test, the proposed CLRHP is compared to MDL, CV, and CVSMAP, respectively. In Fig. 5, the synthesized speech quality in CLRHP is much better than the MDL-based synthesized speech quality. Moreover, the listeners provide higher preference to the proposed CLRHP algorithm compared to the CV and CVSMAP algorithms.

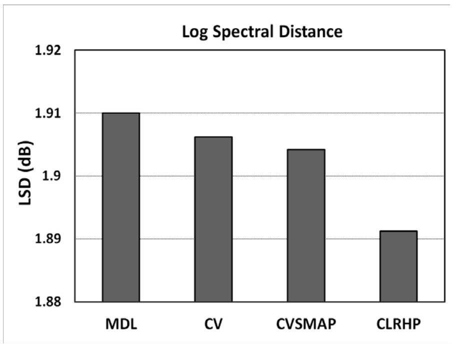

Fig. 4. Log spectrum distance (dB) for each context clustering algorithm.

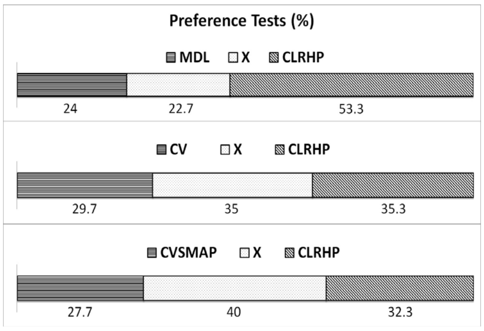

Fig. 5. Scores (%) of preference tests.

V. Conclusions

In this paper, a novel decision tree based context clustering algorithm has been proposed. Unlikely to the conventional context clustering algorithms such as CV and CVSMAP, the proposed algorithm considers the statistical characteristics of the split child nodes. Using the CLRHP, it minimizes the similarity of statistics between each other leaf nodes. The proposed algorithm shows superior performance to conventional context clustering algorithms.

References

1. K. Tokuda, H. Zen, and A.W. Black, “An HMM-based speech synthesis system applied to English,” in Proc. IEEE Speech Synthesis Workshop (2002).
2. K. Shinoda and T. Watanabe, “Acoustic modeling based on the MDL criterion for speech recognition,” in Proc. Eurospeech, 99-102 (1997).
3. K. Shinoda and T. Watanabe, “MDL-based context dependent sub-word modeling for speech recognition,” (in Japanese), J. Acoust. Soc. Jpn, **21,** 79-86, (2000).
4. T. Shinozaki, “HMM state clustering based on efficient cross-validation,” in Proc. ICASSP, 1157-1160 (2006).
5. Y. Zhang, Z. Yan, and F. Soong, “Cross-validation based decision tree clustering for HMM-based TTS,” in Proc. ICASSP, 4602-4605 (2010).
6. H. Zen, and M. Gales, “Decision tree-based context clustering based on cross validation and hierarchical priors,” in Proc. ICASSP, 4560-4563 (2011).
7. K. Shinoda and C.-H. Lee, “A structural Bayes approach to speaker adaptation,” IEEE Trans. SAP, **9**, 276-287 (2001).
8. D. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Communication, **17**, 91-108 (1995).
9. K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.W. Black, and T. Nose, “The HMM-based speech synthesis system (HTS),” http://hts.ics.nitech/ac.jp.
10. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in Proc. of Eurospeech, 2347-2350 (1999).
11. C. Jung, Y. Joo, and H. Kang, “Waveform interpolationbased speech analysis/synthesis for HMM-based TTS systems,” IEEE SP Letters, **12**, 809-812 (2012).

저자 약력

▸ Chi-Sang Jung

Chi-Sang Jung received the B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, Korea, in 2006, 2008, and 2013, respectively.

▸ Hong-Goo Kang

Hong-Goo Kang received the B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, Korea, in 1989, 1991, and 1995, respectively. He was a Senior Member of Technical Staff at AT&T Labs.Research, from 1996 to 2002. In 2002, he joined the School of Electrical and Electronic Engineering, Yonsei University, where he is currently a Professor. His research interests include speech signal processing, array signal processing, and pattern recognition.