Classification of Children with Voice Impairments using Deep Neural Networks

Chien-Lin Huang and Chiori Hori
National Institute of Information and Communications Technology, Kyoto, Japan
E-mail: {chien-lin.huang, chiori.hori}@nict.go.jp Tel:+81-774986316

Abstract— **This paper presents the deep neural networks to classification of children with voice impairments from speech signals. In the analysis of speech signals, 6,373 static acoustic features are extracted from many kinds of low-level-descriptors and functionals. To reduce the variability of extracted features, two-dimensional normalizations are applied to smooth the inter-speaker and inter-feature mismatch using the feature warping approach. Then, the feature selection is used to explore the discriminative and low-dimensional representation based on techniques of principal component analysis and linear discriminant analysis. In such representation, the robust features are obtained by eliminating noise features via subspace projection. Finally, the deep neural networks are adopted to classify the children with voice impairments. We conclude that deep neural networks with the proposed feature normalization and selection can significantly contribute to the robustness of recognition in practical application scenarios. We have achieved an UAR of 60.9% for the four-way diagnosis classification on the development set. This is a relative improvement of 16.2% to the official baseline by using our single system.**

I. INTRODUCTION

Speech is the most preferred and natural modality of communication for human beings [1]. The affective computing ability enables the machine to understand human's emotion [2]. Judgment about voice quality has been mainly subjective and depends on the listener's skill such as speech intelligibility criteria (SIC) [3, 4]. We can help speech pathologists evaluate and monitor voice impairments in children by creating an automatic classification system which can determine the state of the child's phonetic disorder. Many efforts have been devoted to improving the effectiveness of voice and emotion recognition. Acoustic analysis could be a useful tool to diagnose voice diseases. F-ratio and Fisher's discriminant ratio are applied to demonstrate that the detection of voice impairments by performing Mel-frequency cepstral coefficients (MFCCs) [5]. The nonlinear and phase space features are extracted from voice of children with cochlear implantation and hearing aid [3]. The word-based and frame-based features are applied on various classifiers. Then, different fusion techniques are compared to show the performance improvement. Healthy voice has lower correlation dimension in comparison with disordered voice because it have much regularity [6]. This differentiation can be used to separate the disordered voices from healthy voices. It is well know that vocal and voice diseases do not necessarily cause perceptible changes in the acoustic voice signal. Acoustic analysis is a useful tool to diagnose voice diseases being a complementary technique. Neural networks of multilayer perceptron and learning vector quantization are applied to the automatic detection of voice disorders [7]. In addition, the acoustic, linguistic, and semantic information is popular used to detect the speaker status and emotions. For example, the studies in the emotion recognition focus on the prosodic features, in particular pitch, duration and intensity etc. [8, 9]. Moreover, the voice quality features, such as HNR, jitter, shimmer, and MFCC, have been found useful to emotion detection [10]. A multi-modal emotion recognition system is constructed to extract emotion information from both speech and text input. Six emotion types are classified based on 33 acoustic features and emotional keywords [11]. Several classifiers are evaluated for the emotional classification such as Bayes classifier, Gaussian mixture models (GMM), hidden Markov models, decision trees, k-nearest neighbor, and support vector machines [12].

In this study, we propose the deep neural networks (DNNs) with novel methods of feature normalization and selection for the classification of children with voice impairments as shown in Fig. 1. This study proposes two-dimensional (2-D) feature warping approach to normalize the mismatch between speakers and features. Since the feature warping is performed on speaker and feature dimensions, the feature values are mapped into the same range. The benefit of 2-D feature warping is to alleviate the speaker and feature dependency for classification. To compose efficient features, the feature is transformed into a low-dimensional space by using the feature selection of principal component analysis (PCA) and linear discriminant analysis (LDA). PCA is used to project the features to the orthogonal axes. LDA is applied to show distinctive characteristics. According to feature normalization and selection, the deep neural networks are used to classification of children with voice impairments.

This contribution is organized as follows. Section II elucidates the proposed feature analysis techniques. We present deep neural networks for classification of children with voice impairments in Section III. We describe the experimental setup and report a series of experiments in Section IV. Finally, Section V concludes this work.

II. FEATURE ANALYSIS

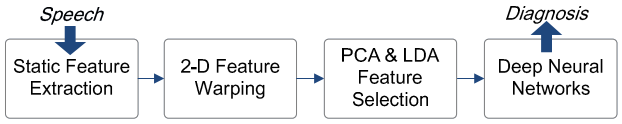

Figure 1. Main diagram of the study.

The feature analysis of this paper is threefold including feature extraction, feature normalization, and feature selection for classification of children with voice impairments from speech.

A. Feature Extraction

Methods of feature extraction can be divided into dynamic and static feature extraction. The dynamic feature extraction is commonly used in the speech and speaker recognition which shows the variation length of feature vectors depended on the duration of audio data [13]–[15]. Instead of the variation length of feature vectors, the static feature extraction is the static length of feature vectors and independent to the duration of audio data. The openSMILE toolkit [16] is used to extract the static feature, in which a 6,373 dimensional feature vector is composed of pitch, energy, zero crossing rate (ZCR), MFCCs, and so on. Functionals are applied to each contour of low-level-descriptors like means, moments, segments, peaks, percentiles, durations, onsets, DCT coefficients, linear and quadratic regression.

B. Feature Normalization

Feature normalization is a technique to smooth variability and to reduce the mismatch problem in speech. The feature warping (FW) provides a transformation mapping from the histogram of each feature vector component to a reference histogram for feature normalization [17]. FW is popularly used in speaker recognition and image processing [18, 19]. Each dimension of feature vector is treated as independent in FW. We estimate the transformation using the cumulative density function (CDF). The density function is defined as Gaussian in this study. Two kinds of variability are variation across speakers and features. In order to remove inter-feature and inter-speaker variability, we proposed the technique of two-dimensional feature warping for feature normalization. The first is the speaker-based feature warping which is used to overcome the problem of speaker dependency. The second feature warping is on the feature level that shows a normalization of each calculated functional feature to the same range.

We estimate the 2-D feature warping on the corresponding $s \times d$ matrix $\mathbf{M}$. d means the dimension of the feature vector. s denotes the number of speaker in the database. Each speech file is identified as a speaker in this study. The speaker-based feature warping is computed by $\hat{\mathbf{M}} = FW(\mathbf{M}(d))$, in which denotes the d-th dimensional feature vector in speech s. The speaker-based feature warping is applied to smooth variety between speakers. After the speaker-based feature warping, the feature-based feature warping $\tilde{\mathbf{M}} = FW(\hat{\mathbf{M}}(s))$ is used to normalize the feature values in speech s. $\mathbf{M}(s) = \mathbf{m}_s = [m_1^s, m_2^s, ..., m_d^s]$ is a high-dimensional feature vector derived from the static feature extraction. With the proposed 2-D feature warping, feature values are bounded in the same range. Note that the order of 2-D feature warping affects the achievable performance in our experiments.

C. Feature Selection

Due to high dimensions (d=6,373) and noise features, the extracted long feature vector is in-effective for the statistical modeling. The feature selection is required in the application. Techniques of PCA and LDA are used for dimensionality reduction via subspace projection to eliminate the noise features. PCA can be realized by using the singular value decomposition (SVD) [20, 21] which finds the optimal projection. SVD is related to the eigenvector decomposition and factor analysis. We perform SVD of the matrix $\tilde{\mathbf{M}}$ as follows:

$$\tilde{\mathbf{M}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular matrix, respectively. T denotes the matrix transposition. Both $\mathbf{U}$ and $\mathbf{V}$ show the orthogonal character. The eigenvector $\hat{\mathbf{U}} = [\hat{u}_1, \hat{u}_2, ..., \hat{u}_R]$ is treated as a transform basis which is empirically selected the first R dimensions. $R \le D$ denotes the projected dimension of the original feature vector in the eigenspace.

PCA is used to project data onto the pairwise linear discriminants and take features as linear combinations of the discriminants [22]. LDA is further applied to minimize the within-class variation and maximize the between-class variation [23]. The solution is obtained through eigenvalue decomposition as follows:

$$S_b v = \lambda S_w v\,. \tag{2}$$

The LDA transform matrix $\mathbf{B}$ is formed as the subset of eigenvectors having the largest eigenvalues λ. S_w and S_b are within-class and between-class scatter matrices, respectively. They are estimated as follows:

$$S_w = \sum_{c=1}^{C}\sum_{i=1}^{N_c}\left(\omega_c^i - \bar{\omega}_c\right)\left(\omega_c^i - \bar{\omega}_c\right)^T, \tag{3}$$

$$S_b = \sum_{c=1}^{C} N_c\left(\bar{\omega}_c - \bar{\omega}\right)\left(\bar{\omega}_c - \bar{\omega}\right)^T, \tag{4}$$

where C means the number of classes that each has the number of N_c feature vectors in the training dataset. $\bar{\omega}_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\omega_c^i$ is the mean of observations in the c-th voice class. ω_c^i indicates the i-th vector in the c-th voice class. $\bar{\omega}$ represents the mean of all instances in the training set.

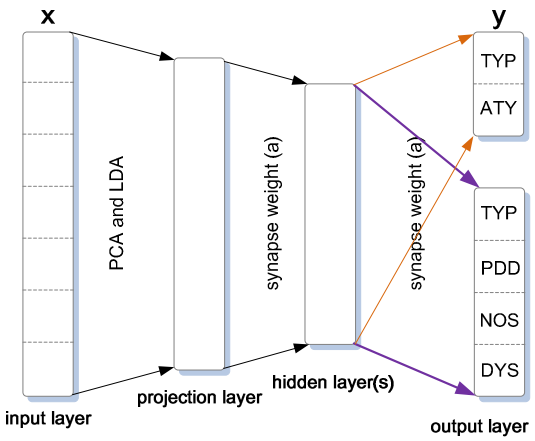

Figure 2. Deep neural networks for voice classification.

III. Deep Neural Networks

After feature analysis, we use the technique of deep neural networks (DNNs) for the classification of children with voice impairments as shown in Fig. 2. The neural networks can be grouped into two categories including feed-forward and recurrent networks according to connection patterns [24]. The difference between feed-forward and recurrent networks is that loops occur in recurrent networks because of feedback connections but no loops are in feed-forward networks. This kind of loops and feedback connections can explain the context information. As a result, the recurrent neural networks are commonly used in language models [25]. We build deep neural networks using the back-propagation learning based on feed-forward architectures with input, hidden and output layers. PCA and LDA are used to transform the input features into projection layer. The values of neurons are estimated as follows:

$$y_j = f(\sum_i x_i \times a_{ji} + b_j) \,, \tag{5}$$

where a_{ji} denotes the synapse weight from node i to node j in the neural network. The variable x_i is un-noise feature which obtained by PCA and LDA. b_j and $f(.)$ are the bias and the activation function, respectively. We use the standard sigmoid function which is the logistic function defined by:

$$f(net_j) = \frac{1}{1+\exp^{-\beta \times net_j}} \,, \tag{6}$$

where β is a slope parameter and $net = \mathbf{x} \cdot \mathbf{A} + \mathbf{b}$. $\mathbf{A}$ is the weight matrix. $\mathbf{b}$ means the set of biases. The continuous sigmoid function is the most frequently used in the neural networks because cumulative distribution functions for many common probability distributions are like sigmoidal. It is a strictly increasing function which exhibits smoothness and has the desired asymptotic properties [24]. The detail parameter setting of DNNs is important for the overall performance. The size of the mini-batches is 200 in this study. The learning rate is defined as 0.6. The cost of negative log-likelihood is used on the output by performing gradient decent. In addition, DNNs are trained with 10 passes through the entire training set.

IV. Experiments

A. Autism Task

The performance is evaluated on INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) [26]. The Autism task features original recordings from children on the autistic spectrum, children with other voice impairments, and a control group. Speech is prompted and covers varying textual content and intonation. We have two sub-tasks: a binary and a four-way classification task. The four-way task "diagnosis" is to classify children into dysphasia, pervasive developmental disorder (PDD), PDD non otherwise specified, or typically developing children by suited voice analysis methods. The binary task is to distinguish typically (TYP) developing children from "atypical" children (ATY), which comprising children with dysphasia (DYS), PDD, or PDD non otherwise specified (NOS) [26].

B. Evaluation Metrics

The classification accuracy is evaluated by unweighted accuracy (UAR) and weighted accuracy on average per class. Since the distribution among classes is not balanced, the competition measure is UAR shown as follows:

$$UAR = \frac{1}{K}\sum_{k=1}^{K}\frac{T_k}{N_k} \times 100\% \,, \tag{7}$$

where K means the total number of class. T_k and N_k are the correct classified number and the total number in the class k. The weighted accuracy (Accuracy) is estimated as follows:

$$Accuracy = \frac{\sum_{k=1}^{K} T_k}{\sum_{k=1}^{K} N_k} \times 100\% \,. \tag{8}$$

We report results based on the classifiers trained per task (like the official baseline) which are better than a single classifier used (multi-task learning) in our experiments.

C. Evaluation of Feature Normalization

In order to know the effect of original features and the proposed feature normalization, we first compare original features with two-dimensional feature warping and other normalizations. The comparison results were illustrated in Table I. The dimension of the features is 6,373. The term "OBS" means the official baseline system [26] using support vector machines (SVM) with the complexity of C=0.01 and C=0.001 for binary and four-way classification, respectively. The linear kernel Support Vector Machines are used on OBS

TABLE I
UAR of the baseline SVM and DNNs system with various normalization methods

Classifier	SVM	DNNs			
Norm	OBS	ORG	LN	MVN	FW
2 classes	92.8%	53.6%	78.8%	93.5%	93.9%
4 classes	52.4%	26.1%	36.4%	54.0%	54.5%

TABLE II
UAR and Accuracy using different DNNs structures

DNNs structure	UAR	Accuracy
Official baseline (SVM)	52.4%	71.2%
6373-100-4	54.5%	76.2%
400-100-4	55.6%	76.8%
400-100-100-100-100-4	60.9%	78.6%

which are known to be robust against overfitting. Instead of SVM or GMM systems [26, 27], we explore DNNs for classification in this study. There are four DNNs systems with different feature normalization. The term of "ORG" indicates the original features without any normalization. Terms of "LN", "MVN", and "FW" mean the length normalization, mean and variance normalization, and the proposed 2-D feature warping, respectively. MVN is a common technique in speech recognition which shows a normalization of each calculated feature vector to a mean of zero and standard deviation of one. The length normalization is used to deal with the non-Gaussian behavior of vectors so that normalized vectors can better fit to the Gaussian assumptions in modeling [28, 29]. Experiments show the significant gap between the original (ORG) and normalized features (LN, MVN, FW) in the DNNs system. Results reflect normalized features make a great impact on DNNs to classify voice impairments. The proposed FW indicates a best UAR. We apply FW on the following experiments.

D. *Evaluation of Feature Selection*

After the 2-D feature warping normalization, we obtain a good speaker and feature invariance feature for the DNNs. We found that DNNs classifiers outperform SVM based on above experiments. Due to the high dimensionality (6,373), it would be a problem such as the high computation and storage cost. We further apply the feature selection of PCA and LDA feature selection on the FW normalized feature. The results are shown in Fig. 3. Figure 3 illustrates the relation between the feature dimension (d) and the percentage of eigenvalue information. We conducted experiments to determine the optimal dimension R in the eigenspace. The eigenvalue reaches 100% information after selecting first 2,600 projected features. We have 99%, 95%, 90% and 80% of eigenvalue information when selecting 2,500, 1,000, 700, and 400 dimensional features. The UAR of the four-way classification task was improved from 54.5% to 55.6% when we only select 400 PCA dimensional (80% eigenvalue information) features.

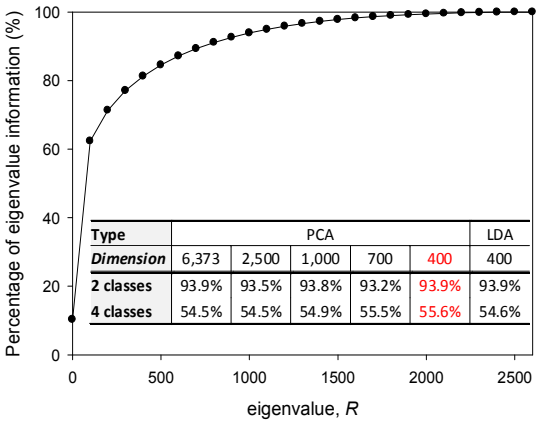

Figure 3. UAR based on the percentage of eigenvalue information for PCA and LDA feature selection.

The results verify the PCA and LDA feature selection effectively extracts important information and reduces redundant components upon the original feature. Especially, PCA offers good results on the four-way classification task.

E. *Evaluation of Network Structures*

We conduct the above experiments by using one hidden layer and 100 neurons. We further explore different network structures and summarize results in Table II. In Table II, the four hidden layers with 100 neurons achieve the best UAR of 60.9% and accuracy of 78.6% in the four-way classification task. In the binary classification task, we found that the neural network of deep layers does not show an apparent superiority to the shallower layers. However, the proposed feature normalization and selection still offer good performances on the binary classification. We achieve the best UAR of 94.6% and accuracy of 93.8% which is based on one hidden layer with 100 neurons (the structure of 400-100-2). In summary, we confirm the proposed DNNs with FW feature normalization and feature selection offer good improvements to classify the children with voice impairments compared with official SVM results.

F. *A Voting Method for the Test Set Results*

Since we have insufficient and unbalanced data in building neural networks, results are sensitive to the initial state and the learning rate. To obtain a robust evaluation on the test set, we use a voting method [30, 31] for combing information of proposed DNNs and SVM [32] systems. On the test set we achieve the best UAR of 92.9% (accuracy of 94.4%) and UAR of 66.0% (accuracy of 82.3%) for binary and four-way classification tasks, respectively.

V. CONCLUSIONS

This paper proposes the novel classification of voice impairments using deep neural networks and two-dimensional feature warping. With the static feature extraction, the two-dimensional feature warping is applied to normalize the variability of speaker and feature coefficients. Then, feature selection is used to produce a lower dimensional and

discriminative components based on techniques of PCA and LDA. We show 16.2% relative improvements over the official baseline on the development set for the four-way diagnosis task. On the test set we achieve the best UAR of 92.9% and 66.0% for binary and four-way classification tasks, respectively. The results are comparable to the official results. Due to insufficient and unbalanced data, it would be interesting to explore stable parameters of DNNs to classify voice impairments in the future.

REFERENCES

[1] T. Lee and P. C. Ching, "Dealing with Imperfections in Human Speech Communication with Advanced Speech Processing Techniques," in *Proc. International Symposium on Signals, Circuits and Systems*, 2011.

[2] R. W. Picard, "Affective Computing," The MIT Press, Cambridge, 1997.

[3] Z. Mahmoudi, S. Rahati, M. M. Ghasemi, V. Asadpour, H. Tayarani, and M. Rajati, "Classification of Voice Disorder in Children with Cochlear Implantation and Hearing Aid using Multiple Classifier Fusion," *Audio, Transactions of the IRE Professional Group*, vol. 30, no. 6, 2011.

[4] Tayarani H, "Effect of Audio-verbal Rehabilitation on Children's Voice Under 12 with Cochlear Implant," MSc thesis Tehran University of Rehabilitation science, College of Speech Rehabilitation, 2002.

[5] J. I. Godino-Llorente, P. Gómez-Vilda and M. Blanco-Velasco, "Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters," *IEEE Trans. Biomedical Engineering*, vol. 53, no. 10, pp. 1943–1953, 2006.

[6] A. Taherkhani, S. A. Seyyedsalehi, A. Mohammadi, and M. H. Moradi, "Nonlinear Signal Processing for Voice Disorder Detection by Using Modified GP Algorithm and Surrogate Data Analysis," *IEEE International Symposium on Signal Processing and Information Technology*, 2007.

[7] J. I. Godino-Llorente and P. Gómez-Vilda, "Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors," *IEEE Trans. Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[8] D. Cairns and J.H. L. Hansen, "Nonlinear Analysis and Detection of Speech under Stressed Conditions," *J. Acoustical Soc. Am.*, vol. 96, no. 6, pp. 3392–3400, 1994.

[9] B. Schuller, M. Wo¨llmer, F. Eyben, and G. Rigoll, *The Role of Prosody in Affective Speech*, pp. 285–307. Peter Lan Publishing Group, 2009.

[10] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," in *Proc. Interspeech*, pp. 2253–2256, 2007.

[11] Z.-J. Chuang and C.-H. Wu, "Multi-Modal Emotion Recognition from Speech and Text" *Computational Linguistic and Chinese Language Processing*, vol. 9, no. 2, pp. 45–62, 2004.

[12] B. Schuller, G. Rigoll, and M. Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture," in *Proc. ICASSP*, pp. 577–580, 2004.

[13] C.-L. Huang and C.-H. Wu, "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis," *IEEE Trans. Computers*, vol. 56, no. 9, pp. 1225–1233, 2007.

[14] C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, "Joint Analysis of Vocal Tract Length and Temporal Information for Robust Speech Recognition," in *Proc. ICASSP*, 2013.

[15] C.-L. Huang and B. Ma, "Maximum Entropy based Data Selection for Speaker Recognition," in *Proc. Interspeech*, 2011.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia*, pp. 25–29, 2010.

[17] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *Proc. 2001: A Speaker Odyssey*, pp.213–218, 2001.

[18] C.-L. Huang, H. Su, B. Ma, and H. Li, "Speaker Characterization Using Long-Term and Temporal Information," in *Proc. Interspeech*, pp. 370–373, 2010.

[19] C. McCool, J. Sanchez-Riera, and S. Marcel, "Feature Distribution Modelling Techniques for 3D Face Verification," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1324–1330, 2010.

[20] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[21] C.-L. Huang, B. Ma, H. Li, and C.-H. Wu, "Speech Indexing Using Semantic Context Inference," in *Proc. Interspeech*, 2011.

[22] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Prentice Hall, Inc., 2001.

[23] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley, New York, 2001.

[24] A. K. Jain and J. Mao, "Artificial Neural Networks: A Tutorial," *IEEE Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[25] S. Kombrink, T. Mikolov, M. Karafi´at, and L. Burget, "Recurrent Neural Network based Language Modeling in Meeting Recognition," in *Proc. Interspeech*, pp. 2877–2880, 2011.

[26] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech*, 2013.

[27] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[28] H. Tang, S. M. Chu, M. Hasegawa-Johnson, and T. S. Huang, "Partially Supervised Speaker Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 959–971, 2012.

[29] C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, "Speaker Clustering Using Vector Representation with Long-Term Feature for Lecture Speech Recognition," in *Proc. ICASSP*, 2013.

[30] D. Jang, M. H. Jin, and C. D. Yoo, "Music Genre Classification Using Novel Features and A Weighted Voting Method," in *Proc. of ICME*, 2008.

[31] K. West and S. Cox, "Features and Classifiers for the Automatic Classification of Musical Audio Signals," *ISMIR04*, 2004.

[32] W. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.