COMBINING MULTIPLE KERNEL MODELS FOR AUTOMATIC INTELLIGIBILITY DETECTION OF PATHOLOGICAL SPEECH

Dong-Yan Huang, Minghui Dong and Haizhou Li

Human Language Technology Department
Institute for Infocomm Research, A*STAR
21-01, Fusionopolis Way, Connexis (South Tower), Singapore 138632

ABSTRACT

Automatic detection of pathological voice is a challenging task in speech processing. Appropriate acoustic cues of voice can be used to differentiate between normal voices and pathological voices. We propose a method to represent each speech utterance using three types of speech signal representations (i.e., cross-correlation matrix, Gaussian distribution and linear subspace) respectively. Various kernels were applied to these representations for measuring resemblance and difference. Four classifiers, i.e., KNN, kernel partial least squares, kernel SVM, and logistic regression, are studied for comparing their performance of classification. Finally, a simple fusion of learning classifiers from different acoustic representations was carried out at the score decision level for enhancing the performance. The different classifiers were evaluated on the Interspeech 2012 challenge development data set and test data set. Their effects in a fusion scheme are studied. The accuracy of the fusion system attained 78.0 % on test set, with an improved gain of 9.1 % over the challenge baseline 68.9 %.

Index Terms— pathological speech, intelligibility, correlation structure feature, multiple kernel models

1. INTRODUCTION

Pathological speech is usually defined as the condition of speech distortion due to functional and organic speech defects and disorders of the organs of speech production [1, 2, 3]. For example, the different locations and the size of the head and neck tumors cause different distortion of speech signals. The defects of any component of speech production (SP) may lead to loss of intelligibility. Speech and voice characteristics of dysarthric speech include imprecise and uncoordinated articulation, variable speech rate, and variation in prosody and rhythm [4]. Intelligibility detection of patients with pathological voice can help possible intervention and rehabilitation. A clinically pathological diagnosis depends mainly on subjective judgement of trained professionals. However, the accuracy of the subjective evaluation is decided by the experience of the listener, which may introduce a range of subjective biases to the diagnostic procedure. Therefore, it is desirable to use noninvasive biomarkers to detect intelligibility of the pathological voice to help the clinician for treatment.

It has been reported that several acoustic cues (source, prosodic and frequency spectrum) are utilized for intelligibility detection [5]. Various machine learning algorithms have been used for intelligibility detection including Gaussian mixture models, neutral network [6], and wavelet modeling [7]. The intelligibility classification of pathological speech has been investigated in the Interspeech pathology sub-challenge [8]. Kim *et al* [9] fused multiple systems to predict intelligibility. Huang *et al* [10, 11] investigated the methods of asymmetric sparse partial least squares without/with kernel for detecting comprehensibility of an utterance. However any interpretation and how the classifier works for intelligibility detection are missed. Recently a mixture of experts approach has been proposed to model complex class boundaries because of heterogeneity in pathological conditions on time-domain analysis of data [12]. The proposed work focused on the exploration of the correlation and covariance of signals to divulge latent variables in the stochastic systems that produce speech and the manifold-based classification algorithms, which analyze data in non-uniform time-frequency domain in terms of lower probability of error and high speed.

In section 2, we present the NKI CCRT database. In section 3, we outline the feature and the correlation feature extraction. We present several principles of classifiers and fusion method in section 4. In section 5, we show the results. Finally, we conclude our study.

2. PATHOLOGICAL DATABASE

The NKI CCRT speech database [13] were used for intelligibility detection. It consists of speech recordings from 55 patients who have neck and head cancer. The intelligibility of their speech were scored on a scale of 1-7. The speech recordings were collected during three stages of Chemo-Radiation Treatment of patients: before Chemo-Radiation Treatment,

Table 1. List of features for intelligibility detection of pathological speech

Features	Statistical functionals
Log HNR, F0, Voicing, Shimmer	mean, range, quartiles std, max, min
Formant frequencies delta MFCCs	
Loudness, RMS Energy, ZCR RASTA-style filtered spectrum	mean, std, max, min avg dur., avg pitch slope

10-weeks after Chemo-Radiation Treatment and 12-months after Chemo-Radiation Treatment. Based on evaluation from thirteen native speech pathologists who speak Dutch, the majority voting rule is employed to obtain scores.

For the purpose of the pathology sub-challenge [8], the binary labels of intelligibility (non-intelligible (NI) and intelligible (I)) were acquired using the midpoint of a scale on same dataset. Overall we have 2385 utterances (NI:1185, I:1200) for the binary classification experiment.

3. FEATURE CONSTRUCTION

3.1. Low Level Features

The study showed the usefulness of voice source (voice quality) and prosodic features for detecting intelligibility, but the spectral features have not shown significant contribution to the final fusion scheme [11]. As the spectral features capture the speech speed, disfluency, and the mispronunciation of phones, we chose delta-mel-cepstra and formant frequencies to manifest alterations in vocal tract shape and dynamics. We hypothesize such alterations that happen with vocal tract aberrations because of tumors. Opensmile is used to extract the features [14] and Table 1 shows the list of features.

Mel-cepstra (MFCCs) are used to generate delta MFCCs by subtracting the 16 mel-cepstra across succeeding data frames with a length of 10 ms in each utterance. The formant dynamics of the vocal tract can be considered as one way to show articulatory alterations in the pathological voice. Formant frequencies can be estimated by a Kalman-based formant predictor [15]. Formant frequencies are computed on data frames with a length of 10 ms. Inspired from a multi-scale approach [16], these two features are extracted at different sample delay spacing 30-ms delay and 10-ms delay, respectively.

3.2. Structure of Feature Correlation

The structure of feature correlation is studied, stimulated by the perception that auto- and cross-correlations of speech signals represent the physiological source changes and coordination of vocal tract trajectories due to a tumor [16].

After extracting speech cues, a set of feature vectors of utterances can be denoted by $F = [f_1, f_2, \cdots, f_n]$, where $f_i \in R^d$ represents the i^{th} utterance with a d-dimensional feature descriptors. Applying to the set of feature vectors, the three types of auto- and cross-correlation, and Gaussian distribution are exploited for their ability of seizing data changes to model pathological voice.

3.3. Multiple Kernel Models

The study has shown that the manifold structure can be exploited for dimensionality reduction or feature extraction as well as improvement of classification performance [17]. Therefore, three types of auto-correlation, covariance matrix, and Gaussian distribution are formulated into manifold structures [18].

3.3.1. Kernels for Linear Space

The auto-correlation of feature set $F = [f_1, f_2, \cdots, f_L]$ can be expressed in a linear subspace $V \in R^{d \times k}$ through singular vector decomposition (SVD) in the following

$$\sum_{i=1}^{L} f_i f_i^T = V \Gamma V^T \tag{1}$$

where L is the number of acoustic features of an utterance, $V = [v_1, v_2, \cdots, v_k]$, v_j is the j^{th} eigenvector, k is the number of columns of the subspace.

The speech samples can be formed an ensemble of linear subspaces, which are considered as the data on Grassmann manifold M (Riemmannian manifold), denoted by $V = V_{i\,i=1}^{N}$, where N is the number of utterances. Using Mercer kernels [19], the similarity can be calculated between two data points V_i and V_j by mapping the Grassmann manifold to Euclidean space. From the principal angles between two subspaces, the projection kernel is given by:

$$K_{i,j}^{proj.-poly.} = (\eta \| V_i^T V_j \|_F^2)^{\alpha} \tag{2}$$

where an element of the kernel matrix is $K_{i,j}^{proj.-poly.}$. The projection function is $\Phi_{proj.} = V_i V_i^T$. A type of RBF kernel can be obtained using $\Phi_{proj.}$ by:

$$K_{i,j}^{proj.-poly.} = (\eta \| \Phi_{proj.}(V_i) - \Phi_{proj.}(V_j) \|_F^2) \tag{3}$$

3.3.2. Kernels for Cross-correlation Matrix

The vocal feature set is also expressed with a cross-correlation matrix of $d \times d$ dimensions:

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (f_i - \hat{f})(f_i - \hat{f})^T \tag{4}$$

where a mean vector of the vocal features is denoted by $\hat{f}$. The nonsingular cross-correlation matrices of $d \times d$ dimensional are symmetric positive definite (SPD) matrices on a Riemmannian manifold. Such non-singular covariance matrices $C = \{C_i\}_{i=1}^{N}$, is considered as data on the symmetric positive definite Riemannian manifold. The distance of log-euclidean (DLE) can be used as distance metric for the SPD matrices. Based on DLE and applying an ordinary matrix logarithm operator to C, a Riemannian kernel was used to compute the inner product of T by mapping the data on the Riemmanian manifold of the SPD matrix to those in the tangent space:

$$K_{i,j}^{dle-poly.} = (\eta \text{trace} \| \log(C_i) \cdot \log(C_j) \|)^{\alpha} \quad (5)$$

The mapping corresponding to $K_{i,j}^{proj.-poly.}$ is obtained by $\Phi_{dle} = \log(C_i)$. As a type of RBF, the kernel can be obtained using Φ_{dle} as

$$K_{i,j}^{dle-rbf} = (\eta \| \Phi_{dle}(C_i) - \Phi_{dle}(C_j) \|_F^2) \quad (6)$$

3.3.3. Gaussian Distribution Kernels

Assuming the feature vectors $f_1, f_2, \cdots, f_n$ with an n-dimensional Gaussian distribution $N(b, \Gamma)$:

$$b = E(f_i) = \frac{1}{n-1} \sum_{i=1}^{n} f_i \quad (7)$$

$$\Gamma = E[(f_i - \mu)(f_i - \mu)^T] = \frac{1}{n-1} \sum_{i=1}^{n} (f_i - \mu)(f_i - \mu)^T \quad (8)$$

where b and Γ are the mean and covariance of feature vectors. They are jointly modeled in a single Gaussian model.

A multivariate Gaussians matrix of d-dimensional is a Riemannian manifold. Incorporated into the space of SPD matrices, a Gaussian matrix $N(b, \Gamma)$ of d-dimensional can be particular expressed by a SPD matrix G of $(d+1) \times (d+1)$ dimensional in the following:

$$N(b, \Gamma) \sim G = |\Gamma|^{-\frac{1}{d+1}} \begin{bmatrix} \Gamma + bb^T & b \\ b^T & 1 \end{bmatrix} \quad (9)$$

Using the SPD matrices $\mathcal{G} = G_{i}{}_{i=1}^{N}$, the corresponding Riemannian kernels can be calculated as:

$$\begin{array}{rcl} K_{i,j}^{dle-poly.} & = & (\eta \text{trace} \| \log(G_i) \cdot \log(G_j) \|)^{\alpha} \\ K_{i,j}^{dle-rbf} & = & \exp(\eta \| \Phi_{dle}(G_i) - \Phi_{dle}(G_j) \|_F^2) \end{array} \quad (10)$$

4. COMBINATION OF CLASSIFIERS

The feature construction approach may produce multiple features from different domains. Moreover, the patterns of features and correlation between features may be different for each subject. Thus, it is desirable to develop a multivariate fusion and classifiers that can efficaciously combine both without and with contextual information from multiple input features and subject identity (e.g., gender) in making its pathological prediction.

4.1. Classifiers

In [9], K-nearest neighbor (KNN) was used to assess the discriminating capacity of the feature sets at sentence-level of individual sub-systems and feature fusion in comprehensibility detection using the NKI CCRT Speech Corpus. The KNN shows promising results for intelligibility classification. Therefore, KNN is used to evaluate the features from different domains. For this work, the kernel SVM in a LibSVM [20] implementation is employed on the pre-computed Riemannian kernel matrices for classification. A logistic regression with L_2-regularization is applied to these sets of features for detection. The Liblinear implementation is exploited for optimization [21]. The asymmetric partial least squares classifier is used for classification as well [11].

4.2. Combination System

Individual classifier learns six Riemannian kernels with different speech features. A term w is introduced for score fusion of different classifiers at decision level:

$$S^{fusion} = \sum_{i=1}^{M} w_i S_i \quad (11)$$

M is the number of classifiers, w_i is the weight for the i-th classifier, and S_i is the score of i-th classifier.

5. EXPERIMENTAL RESULTS

We apply the four classifiers (KNN, kernel SVM, LR, and ASIMPLS) to the different aspects of speech production in Table 1: speech source cues, vocal tract cues and prosodic cues. Although the weighted average recall is also presented in this study, the unweighted average recall of classes I and NI is used to compare all classification performances with the challenge dataset in order to align with the challenge setup.

Table 2 shows the performance of four classifiers learned from six Riemannian kernels using different speech features of the NCSC test sets. The spectral feature set of formant frequencies and delta MFCCs at multi-scales has the highest unweighted average recall of the individual feature set. It implies that the correlation of this spectral feature set can capture the changes that occur with the vocal tract shape and dynamics due to the location and size of tumors.

Considering the high scores in subsystems such as Logistic Regression with kernel (proj.-poly.) on voice source

Table 2. *Unweighted accuracy recall (%) on the test set based on different feature sets*

(a) Voice Source Cues

	Linear Space		Cross-correlation matrix		Gaussian Matrix	
	proj.-poly. kernel	proj.-rbf kernel	dle-poly. kernel	dle-rbf kernel	dle-poly. kernel	dle-rbf kernel
KNN	66.10 %	65.89 %	66.02 %	54.23 %	63.84 %	60.26 %
SVM	64.56 %	63.27 %	64.47 %	52.34 %	61.78 %	58.24 %
Logistic Regression	71.64 %	61.04 %	44.07%	62.87%	42.78 %	60.76 %
ASIMPLS	69.40 %	69.42 %	69.47 %	69.52 %	61.78 %	68.85 %

(b) Vocal Tract Cues

	Linear Space		Cross-correlation matrix		Gaussian Matrix	
	proj.-poly. kernel	proj.-rbf kernel	dle-poly. kernel	dle-rbf kernel	dle-poly. kernel	dle-rbf kernel
KNN	68.23 %	65.89 %	69.42 %	53.67 %	68.84 %	67.63 %
SVM	69.48 %	66.72 %	70.47 %	56.74 %	69.78 %	66.67 %
Logistic Regression	69.45 %	68.35 %	48.73%	61.78%	48.96 %	68.36 %
ASIMPLS	73.10 %	73.12 %	70.91 %	70.42 %	69.58 %	70.65 %

(c) Prosody Cues

	Linear Space		Cross-correlation matrix		Gaussian Matrix	
	proj.-poly. kernel	proj.-rbf kernel	dle-poly. kernel	dle-rbf kernel	dle-poly. kernel	dle-rbf kernel
KNN	68.89 %	66.82 %	68.62 %	58.23 %	66.04 %	64.26 %
SVM	70.56 %	68.65 %	44.27 %	69.34 %	46.78 %	67.24 %
Logistic Regression	70.94 %	68.04 %	44.07%	69.87%	45.78 %	67.67 %
ASIMPLS	72.02 %	72.02 %	70.17 %	71.82 %	70.78 %	71.95 %

features, ASIMPLS with kernel (proj.-rbf) on vocal tract features, and ASIMPLS with kernel (proj.-poly.) on prosody features, we propose a system using the fusion of these three subsystems.

Then a comparative study was performed on the proposed system, the winner system [9], the 2nd place system [22], the 3rd place system [10], ASKPLS [11], the baseline RF [8], and the baseline SVM [8]. While the accuracy of classification on the development set manifests 74.1 % of the unweighted average recall, the test set classification provides 78 % accuracy in Table 3.

Table 3. *The comparative results of the proposed method, the winner system [9], the 2nd place system [22], the 3rd place system [10], ASKPLS [11], the baseline SVM, and the baseline RF [8]).*

System	dev set (%)	test set (%)
Proposed method	74.1	78.0
Joint classification([9])	79.9	76.8
S-GPR+KPCA ([22])	77.6	73.7
ASPLS ([10])	66.0	71.9
ASKPLS ([11])	62.9	74.0
Baseline RF ([8])	64.8	68.9
Baseline SVM ([8])	61.1	68.0

6. CONCLUSION

We propose to model each speech utterance using three types of speech signal representations (i.e., linear subspace, Gaussian distribution, and covariance matrix) respectively. Different kernels are applied to these representations for measuring similarity and difference. We explored the discrimination capabilities of each representation for detecting the intelligibility of pathological voice. The correlation structure of formant frequencies and the delta MFCCs at multi-delay scales have more capabilities in discrimination of pathological voice from normal voice. For the test set, the final proposed system can achieve an unweighted average recall performance of 78.0%. We continue to investigate more salient speech cues for intelligibility detection of the pathological voice in the future.

7. REFERENCES

[1] S Hadjitodorov and P Mitev, "A computer system for acoustic analysis of pathological voices and laryngeal diseases screening," *Medical Engineering & Physics*, vol. 24, no. 6, pp. 419–429, 2002.

[2] R Jones, "Observations on stammering after localized cerebral injury," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 29, no. 3, pp. 192, 1966.

[3] C Lai, S Fisher, J Hurst, F Vargha-Khadem, and A Monaco, "A forkhead-domain gene is mutated in a severe speech and language disorder," *Nature*, vol. 413, no. 6855, pp. 519–523, 2001.

[4] R. D. Kent, "Intelligibility in speech disorders: theory, measurement and management," *Journal of the ACM (JACM)*, vol. 1, 1992.

[5] A. Dibazar, S. Narayanan, and T. Berger, "Feature analysis for automatic detection of pathological speech," in *in Proc. of IEEE EMU'S meeting*, 2002.

[6] R. Ritchings, M. McGillion, and C. Moore, "Pathological voice quality assessment using artificial neural networks," *Medical Engineering & Physics*, vol. 24, no. 7, pp. 561–564, 2002.

[7] R. Behroozmand and F. Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients speech signal with unilateral vocal fold paralysis," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 474–485, 2007.

[8] B. Schuller, S. Steidl, A. Batliner, E Nöth, A. Vinciarelli, F. Burkhardi, R. van Son, F. Wehinger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 2012, ISCA.

[9] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S.S. Narayanan, "Intelligibility Classification of Pathological Speech Using Fusion of Multiple Subsystems," in *13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 2012, ISCA.

[10] D.-Y. Huang, Y. Zhu, D. Wu, and R. Yu, "Detecting Intelligibility by Linear Dimensionality Reduction and Normalized Voice Quality Hierarchical Features," in *13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 2012, ISCA.

[11] D.-Y. Huang, M. Dong, and H. Li, "Intelligibility detection of pathological speech using asymmetric sparse kernel partial least squares classifier," in *Proceedings of IEEE Conference on Acoustic, Speech and Signal Processing*. IEEE, 2014, pp. 3744–3748.

[12] R. Gupta, K. Audhkhast, and S. Narayanan, "A Mixture Of Experts Approach Towards Intelligibility Classification of Pathological Speech," in *Proceedings of IEEE Conference on Acoustic, Speech and Signal Processing*. IEEE, 2015, pp. 1986–1990.

[13] L. van der Molen, M. van Rossum, A. Ackerstaff, L. Smeele, C. Rasch, and F. Hilgers, "Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients / views," *BMC Ear Nose Throat Disorders*, vol. 9, no. 10, 2009.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *ACM Multimedia*. ACM, 2010, pp. 1459–1462.

[15] D. Mehta, D. Rudoy, and P. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, 2012.

[16] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal Biomarkers of Depression Based on Motor Incoordination," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, Florence, Italy, 2013, ACM, pp. 41–48.

[17] M. A. Davenport, C. Hegde, M. B. Wakin, and R. G. Baraniuk, "MANIFOLD-BASED APPROACHES FOR IMPROVED CLASSIFICATION," 2007.

[18] M. M Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X Chen, "Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild," in *International Conference on Multimodal Interaction*. ACM, 2014, pp. 494–501.

[19] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *ICML*. 2008, ACM.

[20] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[21] R.-E. Fan, K.-W. Chang, X.-R. Wang C.-J. Hsieh, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[22] D. Lu and F. Sha, "Predicting Likability of Speakers with Gaussian Processes," in *13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 2012, ISCA.