INTELLIGIBILITY DETECTION OF PATHOLOGICAL SPEECH USING ASYMMETRIC SPARSE KERNEL PARTIAL LEAST SQUARES CLASSIFIER

Dong-Yan Huang, Minghui Dong and Haizhou Li

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

ABSTRACT

Pathological speech usually refers to the voice disorders resulting from atypicalities in voice and/or in the articulatory mechanisms due to disease, illness or other physical problem in the speech production system. It may increase unhealthy social behavior and voice abuse, and dramatically affect the patients' quality of life. Therefore, automatic intelligibility detection of pathological speech has an important role in the opportune treatment of pathological voices. This paper proposes to use asymmetric sparse kernel partial least squares classifier (ASKPLSC) for intelligibility detection of pathological speech. The proposed approach achieves an unweighted accuracy (UA) of 74.0%, which is 7.34% relative improvement of baseline system of an UA of 68.90% for the Pathology Sub-Challenge of INTERSPEECH 2012 Speaker Trait Challenge.

Index Terms— Pathological speech, intelligibility of speech, kernel function, sparse kernel partial least squares regression, asymmetric sparse kernel partial least squares classifier

1. INTRODUCTION

Verbal communication is important for our daily life. However, the quality of speech communication can be affected by the speech disorders, which can be the consequences of a variety of causes on the vocal production system. For example, head and neck cancers may cause the difficulties of speech understanding in the perception by others, and surgical treatment and other inventions can even result in a degenerating effect on patients' production organ function, e.g. the quality of voice and speech. As there are many factors causing the pathological speech, their diagnostic assessment, choosing an appropriate treatment, and monitoring have attracted increasing attention in clinical domain. The pathological speech assessment is mainly based on the subjective judgment of trained professionals. This subjective evaluation suffers from dependency on the experience of the listener and on its consistency on judging pathological voice quality. Therefore, in clinical practice there is a great demand for objective and automatic methods for assessing and classifying pathological speech quality. It is desirable to provide the assessment medical devices featuring high accuracy, reliability, scalability and low cost of processing.

Previous research and studies have shown that a number of acoustic features such like voice quality features, prosodic features, phonemic features, perceptual features, automatic speech recognition based features have discriminative power for intelligibility classification of pathological speech [1, 2, 3]. They have been used for the automatic assessment of intelligibility for pathological speech. Although a large number of features have been developed for intelligibility assessment, the problem is still challenging because of the wide variability of characteristics of disordered speech. The relationship between the perceptual intelligibility and the abnormal variation of disordered speech is still an on-going study. The wide variability in speaker factors such like gender, age, native/non-native, dialectal, and the atypicalities in voices, makes the problem even tough. These issues are surely encountered in development of automatic intelligibility assessment system working in real world scenarios.

This paper aims to design intelligibility classification system using the database provided by the organizers for the Pathology Sub-Challenge (PSC) of Interspeech 2012 Speaker Trait Challenge [4]. The database consists of sentence-level speech in Dutch spoken by patients having head and neck cancer prior to and after concomitant chemo-radiation treatment [5]. Various location and size of tumors may have a detrimental effect on speech production mechanism, even resulting in intelligibility loss of the speech. Due to these physical properties of disordered speech, the acoustic features expose non-linearities. A few prosodic features have been shown to possess significant discriminative power in classifying intelligibility of pathological speech [2].

Among previous automatic intelligibility assessment systems, a semi-supervised sparse Gaussian processes has been proposed for modeling data [6], which showed higher accuracy classification than the baseline. They proposed a novel algorithm to greedily select a sparse subset of features and then use kernel PCA to compute a joint embedding of the training, development and test instances. Different from their approach, we propose asymmetric sparse kernel partial least squares approach for the prediction task. This novel approach

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

can achieve a performance of unweighted accuracy of 74.0 % on intelligibility detection at much lower cost processing time than the system of UA of 73.70 % [6].

This paper is organized as follows. In Section 2, we give a brief explanation about the baseline acoustic features and systems of the challenge, we explain each subsystem. Next, we present in detail asymmetric sparse kernel partial least squares classifier. Then, we show the experimental results and discuss the whole system. Finally, we conclude our study and provide directions for future work.

2. ACOUSTIC FEATURES

For the Pathology Sub-Challenge, the "NKI CCRT Speech Corpus" is used for intelligibility detection. This corpus consists of pathological Dutch speech spoken by patients prior to and after concomitant chemoradiotherapy (CCRT) due to the cancer of the head and neck. The baseline features are in total number of 6125 and comprise low-level descriptors (LLDs), their first-order derivatives, and the functionals applied thereon. LLDs include energy, spectral and voicing related features. The baseline systems are a SVM and a random forest applying to those baseline features. We would like to study whether these feature sets are suitable well to our proposed approach for the automatic assessment of intelligibility for pathological speech.

2.1. Prosodic and intonational features

In the paper [7], the authors found that no-intelligible (NI) speakers have difficulty in pronouncing a few specific speech sounds, resulting in abnormal prosodic and intonational shapes. The F0 trajectory is not smooth for NI speakers. The phoneme and utterance level acoustic features have been proposed to distinguish the differences between I and NI speakers. Inspired by their work, we combine the features such like the F0 features, frame-based energy envelop, the spectral envelop, and the sum of auditory spectrum for describing the prosodic and intonation of speech. The discriminative power of these features for I/NI classes is tested by our proposed approach.

2.2. Voice quality features

It has been reported that laryngeal tumors can impede vocal fold movements, causing voice quality distortion [8]. These distortions (breathy, tense, hoarse, or modal voice) decreases dramatically the intelligibility. In our system, the voice quality features comprise logarithmic harmonics-to-noise ratio (HNR), loudness, RMS energy, sum of the RelAtive SpecTrAl (RASTA) style filtered auditory spectrum, zero-crossing-rate (ZCR), pitch by SHS + Viterbi smoothing, probability of voicing, as well as voice quality features such as Shimmer (Local). The harmonicity features attempt to capture the various form of periodicity disturbances in the acoustic signal. The pathological voices are characterized by loss of harmonicity. In order to capture the dynamics of a signal in terms of some relatively static acoustic parameters, the delta features, which estimate first derivatives of each feature dimension as a way of converting dynamic and contextual behavior into a relatively fixed point in the new feature space, are added into the set. The individual acoustic feature for discriminating the pathological voices has been investigated in various medical literature [6, 9, 10]. From the speech production point of view, these acoustic features should not be independent. The motivation of combination of aperiodicity features and noise features as a group is to understand the impact of the mutual information of acoustic features on the discrimination between normal and pathological voices.

2.3. Spectral features

The spectral features are the most widely used features in speech processing such as mel-frequency cepstral coefficients (MFCCs), the RASTA style filtered auditory spectra and spectral. In fact, some studies have adapted the techniques of Automatic Speaker Recognition (ASR) to pathological voice assessment [11]. Previous studies have shown that non-laryngeal tumors in the vocal tract can have a negative influence on articulation in speech production, for instance a shift in localization of articulation, modified articulatory tension and compensatory articulation [12]. These cause the changes in the spectral flux, spectral energy, psychoacoustic sharpness. In stead of selecting the traditional spectral features such as MFCCs, the spectral energy, flux, variance, Skewness, Kurtosis, slope, and their functionals are used for the assessment of intelligibility. These features capture the rate of speech, disfluency, and the mispronunciation of phones.

3. ASYMMETRIC SPARSE KERNEL PARTIAL LEAST SQUARES CLASSIFIER (ASKPLSC)

As the acoustic features expose non-linearities due to the physical properties of disordered speech, we apply a kernel function for modelling these non-linearities and asymmetric sparse kernel partial least squares classifier for the automatic intelligibility detection of the pathological speech.

3.1. Kernel Transformation

The kernel transform is achieved by calculating the similarity between elements for a finite set of features $s_n, i = 1, 2, \cdots, N$. There are many alternatives for kernel function like polynomial kernel, Gaussian kernel or sigmoid kernel. The covariance matrix of Gaussian K whose ij-th element is define as

$$K(s_i, s_j) = \exp \frac{-\|s_i - s_j\|^2}{2l^2} \quad (1)$$

where l is the width of a scale for the kernel. The selection of l is not highly crucial, usually it is enough to find a decent range for l. In this paper, we use the Gaussian kernel since it is simple and efficient.

3.2. Kernel Normalization

To force the bias term of the conversion model to zero, kernel centering is used. Centering in the kernel space is not as obvious as in the original feature space, since the mean cannot be computed directly. In this paper, for training and testing kernel $K(s_i, s_j)$, we apply mean and variance normalization (MVN) which transforms every column and row vectors of the feature matrix to a random variable with zero mean and unit variance.

3.3. Sparse Kernel Partial Least Squares Regression

Partial Least Squares (PLS), which aims to model the relationship between two data sets (blocks of variables) by means of *score matrix* (latent matrix), has attracted much attention for handling high-dimensionality and multicollinearity in several areas of scientific research. However, PSL cannot automatically select the relevant variables during dimension reduction step [13]. Due to this fact, Sparse Partial Least Squares is proposed by imposing sparsity on the dimension reduction step of PLS, so that it might be possible for one to reduce dimension and select distinctive variables simultaneously. Therefore, we propose a sparse kernel partial least squares regression in the following for a normalized kernel matrix $\mathbf{K}$.

Kernel Partial Least Squares (KPLS) aims to decompose the *response matrix* ($\mathbf{Y} \in \mathbf{R}^{n\times q}$) and the *predictor kernel matrix* ($\mathbf{K} \in \mathbf{R}^{n\times p}$) into the forms

$$\begin{aligned} \mathbf{Y} &= \mathbf{TQ}^T + \mathbf{F} \\ \mathbf{K} &= \mathbf{TP}^T + \mathbf{E}, \end{aligned} \tag{2}$$

where $\mathbf{T} \in \mathbf{R}^{n\times k}$, called score matrix, produces k linear scores; $\mathbf{Q} \in \mathbf{R}^{q\times k}$ and $\mathbf{P} \in \mathbf{R}^{p\times k}$ are the matrices of coefficients (loadings); $\mathbf{F} \in \mathbf{R}^{n\times q}$ and $\mathbf{E} \in \mathbf{R}^{n\times p}$ are the matrices of random errors. For the purpose of effective decomposition, KPLS turns to seek $\mathbf{K}$ direction vectors $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_k)$ such that $\mathbf{T} = \mathbf{KW}$, instead of seeking $\mathbf{T}$ directly. To find the i-th direction vectors $\mathbf{w}_i$, the criterion for univariate $\mathbf{Y}$ is formulated as

$$\begin{aligned} \mathbf{w}_i = &\arg\min_{w} \text{cor}^2(\mathbf{Y}, \mathbf{K}w)\text{var}(\mathbf{Kw}) \\ &\text{s.t. } \mathbf{w}^T\mathbf{w} = 1 \text{ and } \mathbf{w}^T\Sigma_{KK}\mathbf{w}_j = 0, \end{aligned} \tag{3}$$

for $j = 1, \cdots, i-1$, where $\Sigma_{\mathbf{KK}}$ is the covariance of $\mathbf{K}$, and $\sigma_{\mathbf{KY}}$ is the covariance of $\mathbf{K}$ and $\mathbf{Y}$. Note that $\mathbf{w}^T\sigma_{\mathbf{KY}}{\sigma_{\mathbf{KY}}}^T\mathbf{w} = \sum_{j=1}^{q}\text{cov}^2(\mathbf{Kw}, \mathbf{Y}_j)$, thus the criterion (3) is also suitable for univariate KPLS. Consequently, the criterion is reformulated as

$$\begin{aligned} \hat{\mathbf{w}}_i = &\arg\min_{\mathbf{w}} \mathbf{w}^T\mathbf{K}^T\mathbf{Y}\mathbf{Y}^T\mathbf{Kw} \\ &\text{s.t. } \mathbf{w}^T\mathbf{w} = 1 \text{ and } \mathbf{w}^T S_{\mathbf{KK}}\mathbf{w}_j = 0, \end{aligned} \tag{4}$$

for $j = 1, \cdots, i-1$. Once obtaining the latent scores $\mathbf{T}$, loading $\mathbf{Q}$ is estimated for the model $\mathbf{Y} = \mathbf{TQ}^T + \mathbf{F}$, and $\hat{\beta}^{PLS} = \hat{\mathbf{W}}\hat{\mathbf{Q}}^T$, where $\hat{\mathbf{W}}$ and $\hat{\mathbf{Q}}$ are the estimations of $\mathbf{W}$ and $\mathbf{Q}$. Therefore, $\mathbf{Y} = \mathbf{TQ}^T + \mathbf{F} = \mathbf{KWQ}^T + \mathbf{F} = \mathbf{K}\hat{\beta}^{PLS} + \mathbf{F}$.

In conclusion, KPLS utilizes a small number of latent scores based on a basic latent decomposition to linearly regress the original variables, so as to avoid multicollinearity. KPLS estimator $\hat{\beta}^{PLS}$ is no longer consistent, as well as hardly interpretable, when facing a very large p and small n in complete generality [13]. For consistent estimator and interpretable direction vectors, sparsity is imposed into dimension reduction step of KPLS.

Directly imposing the L_1 constraints to the original direction vectors a tends to be not sparse enough, and the problem is not convex. Instead, the SKPLS criterion is presented as a generalized regression formulation by imposing L_1 penalty into a surrogate of direction vector c, which is close to a:

$$\begin{aligned} &\min_{\mathbf{a},\mathbf{c}} -\kappa\mathbf{a}^T\mathbf{Ma} + (1-\kappa)(\mathbf{c}-\mathbf{a})^T\mathbf{M}(\mathbf{c}-\mathbf{a}) + \lambda_1|\mathbf{c}|_1 + \lambda_2|\mathbf{c}|_2 \\ &\quad\text{s.t. } \mathbf{a}^T\mathbf{a} = 1 \end{aligned} \tag{5}$$

where $\mathbf{M} = \mathbf{K}^T\mathbf{YY}^T\mathbf{K}$. Note that the first L_1 penalty guarantees the sparsity of $\mathbf{c}$, while the second L_2 penalty keeps the potential singularity in $\mathbf{M}$ when solving $\mathbf{c}$. Letting $\mathbf{c}_1, \cdots, \mathbf{c}_q \in \mathbf{R}^p$ denote the sparse surrogate direction vectors resulting from the SKPLS method, a classification rule and a bias can be obtained by performing standard LDA on the matrix $(\mathbf{Kc}_1, \cdots, \mathbf{Kc}_q)$.

3.4. SKPLS-based Asymmetric Pattern Classification

The SKPLS regression is able to classify the samples to close to their own center. According to our observation, the boundary of two classes always goes through the point (0,0). For unbalanced binary classes, classification hyper plane can be further generated based on the method proposed in [14] to compensate the loss of the larger class dataset.

4. EXPERIMENTAL RESULTS

As an ASPLS-based classifier has been proposed for pathological speech detection [15], we would like to compare the performance of the proposed classifier with it as well as the state-of-the-art technologies.

An extensive acoustic feature set is provided by the organizers, which is extracted with openSMILE. The 6125-dimensional feature vectors are divided into 13 groups shown in Table 1 and employed our classifier to optimize performance of each. These groups are used to define different base Gaussian RBF kernels. We used speaker mean normalization on these feature sets to remove the effect of speaker variability. We noticed that the bias of optimal vectors is smaller than the non speaker normalization for all the subsets. As the speech in NCSC was recorded in different session, the session variability caused by different channel affects the system performance. Eigen-channel factor estimation technique has been previously proposed to solve a similar problem by minimizing the effect of speaker state channel variability [16]. As PLS aims at iteratively finding the dominant eigenvectors of the problem (4), the effect of session variability can be removed by PLS (KPLS, SPLS, SKPLS) regression.

The unweighted accuracies (UA)s are computed with ASKPLSC on training data. From Table 1, we observe that the features such as spectral, f0, log HNR, and Shimmer (local) appear to have significant discriminative power for pathological speech. All of them increase almost 7-10 % on the unweighted accuracy for the test data. As mentioned in section 2, NI speakers have difficulty to pronounce a few vowel sounds. The f0 features can capture the mispronunciation of phones. The pathological voices are characterized by loss of harmonicity. The features such as log HNR, Shimmer (local), and spectral can represent well the harmonicity, the rate of speech, and disfluency. The results obtained by model trained by our proposed approach reveals these facts. It implies that our proposed modelling approach is suitable to automatically assess the intelligibility of pathological speech. Moreover, these models are flexible as each base kernel has its own parameters to be tuned.

Table 1. *13 features groups: names of the low-level descriptors (LLDs), individual unweighted accuracy recall for the pathology development and test data.*

LLD Names	Grp no	D	UA$_d$	UA$_t$
Energy Related Groups				
Sum of aud. sp.	1	96	63.34	66.46
Sum of RASTA aud.	2	96	55.63	58.67
RMS energy	4	96	61.03	64.05
Zero-crossing rate	5	96	60.01	60.60
Spectral Related Groups				
RASTA-style aud. sp.	3	2496	54.85	60.49
MFCC 1-14	7	1344	55.63	58.67
spectral	6	1344	60.78	70.26
Voicing Related Groups				
F0	8	92	54.34	65.96
Voicing	9	92	60.76	63.67
Jitter local	10	92	61.03	64.05
Jitter delta	11	92	61.39	62.69
Shimmer local	12	92	61.44	68.87
Log HNR	13	92	61.34	68.89

Then we conduct a comparison study on the baseline SVM [4], the baseline RF [4], the 1st Prize system of the PSC [7], the 2nd ranking system of the PSC [6], the 3rd ranking system of the PSC [15], and ASKPLS system using the fusion of feature groups (1,2,4,5,8,9,12,13).

From Table 2, the ASKPLS classifier outperforms the baseline systems, the ASPLS system and the 2nd ranking system of the PSC. While classification on the develop set shows the unweighted average recall of 74% on test data (5.1 % absolute gain) over the baseline system (68.9 %). The performance of the proposed system is still lower than the 1st prize system of the PSC. We shall develop new machine learning algorithms for automatic assessment of the pathological speech in the future. In Table 2, ”dev set” and ”test set” indicate the classification accuracies tested on development set and test set, respectively. Note that the baseline systems and our systems are tuned on development set and tested on development set or test set, while the system [7] are tuned on train set and tested on development set, or tuned on development set and tested on test data. The system [6] is 5-fold CV average accuracy for ”development set”, or tuned on development set and tested on test data.

Table 2. *The comparison study on the baseline SVM, the baseline RF [4]), the 1st Prize system of the PSC [7], the 2nd ranking system of the PSC [6], the 3rd ranking system of the PSC [15], and ASKPLS system using the fusion of feature groups (1,2,4,5,8,9,12,13).*

System	dev set (%)	test set (%)
Baseline SVM ([4])	61.1	68.0
Baseline RF ([4])	64.8	68.9
Joint classification([7])	79.9	76.8
S-GPR+KPCA ([6])	77.6	73.7
ASPLS ([15])	66.0	71.9
KASPLS (1,2,4,5,8,9,12,13)	62.9	74.0

5. CONCLUSION

Automatic intelligibility detection of pathological speech has significant values in real-world applications in the medical domain. We proposed ASKPLSC to evaluate the discrimination capabilities of several acoustical feature sets for detecting the intelligibility of pathological speeches. The results showed that the modellings obtained by the proposed approach can detect the intelligibility of pathological speech, which match well the diagnostic assessment by experts. The system features fast and achieves an unweighted average recall performance of 74 % on the test set, an improvement of 5.1 % absolute gain over the reported baseline system accuracy of 68.90%. We continue to develop classifiers with high prediction accuracy for the automatic assessment of the pathological speech in the future.

6. REFERENCES

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

[2] A. Maier, T. Haderlein, F. Stelzle, E. Noth, E. Nkenke, F. Rosanowski, A. Schutzenberger, and M. Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," *EURASIP J. Audio Speech Music Process*, vol. 2010:926951, 2010.

[3] G. Van Nuffelen, C. Middag, M. De Bodt, and J.P. Martens, "Speech Technology-Based Assessment of Phoneme Intelligibility in Dysarthria," *International Journal of Language and Communication Disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] R. Kazi, R. Venkitaraman, C. Johnson, V. Prasad, P. Clarke, P. Rhys-Evans, C. M. Nutting, and K. J. Harrington, "Electroglottographic comparison of voice outcomes in patients with advanced laryngo-pharyngeal cancer treated by chemoradiotherapy or total laryngectomy," *Int J Radiat Oncol Biol Phys*, vol. 70:344352, 2008.

[9] P. Lieberman, "Pertubation in vocal pitch," *J. Acoust. Soc. Am.*, vol. 33, 1961.

[10] D. Michaelis, M. Frohlich, and H. W. Strube, "Selection and combination of acoustic features for the description of pathologic voices," in *J. Acoust. Soc. Am.*, 1998, vol. 103, pp. 1628–1639.

[11] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Goovanni, and A. Ghio, "Application of automatic speaker recognition techniques to pathological voice assessment (dys-phonia)," in *Proceedings of the EuroSpeech'05*, Lisbon, Portugal, September 2005, pp. 149–152.

[12] I. Jacobi, L. van der Molen, H. Huiskens, M. A. van Rossum, and F. J. M. Hilgers, "Voice and speech outcomes of chemoradiation for advanced head and neck cancer: a systematic review," *Eur Arch Otorhinolaryngol*, vol. 267(10):14951505, October 2010.

[13] H. Chun and S. Keles, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of the Royal Statistical Society - Series B*, vol. 72, no. 1, pp. 3–25, 2010.

[14] D.Y. Huang, S.S. Ge, and Z. Zhang, "Speaker state classification based on fusion of asymmetric simpls and support vector machines," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[15] 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.

[16] M. Li, A. Metallinou, D. Bone, and S. Narayanan, "Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2012.