Automatic intelligibility assessment of pathologic speech in head and neck cancer based on auditory-inspired spectro-temporal modulations

Xinhui Zhou[1], Daniel Garcia-Romero[1], Nima Mesgarani[1], Maureen Stone[2], Carol Espy-Wilson[1], Shihab Shamma[1]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, USA
[2]Departments of Neural and Pain Sciences and Orthodontics, University of Maryland Dental School, Baltimore, USA

[1]{zxinhui, dgromero, mnima, espy, sas@umd.edu}, [2]mstone@umaryland.edu

Abstract

Oral, head and neck cancer represents 3% of all cancers in the United States and is the 6th most common cancer worldwide. Depending on the tumor size, location and staging, patients are treated by radical surgery, radiology, chemotherapy or a combination of those treatments. As a result, their anatomical structures for speech are impaired and this leads to some negative impact on their speech intelligibility. As a part of the INTERSPEECH 2012 speaker trait Pathology sub-challenge, this study explored the use of auditory-inspired spectro-temporal modulation features for automatic speech intelligibility assessment of those pathologic speech. The averaged spectro-temporal modulations of speech considered as either intelligible or non-intelligible in the challenge database were analyzed and it was found that the non-intelligible speech tends to have its modulation amplitude peaks shift towards a smaller rate and scale. Based on SVM and GMM, variants of spectro-temporal modulation features were tested on the speaker trait challenge problem and the resulting performances on both the development and the test datasets are comparable to the baseline performance.

Index Terms: Oral, head and neck cancer, speech pathology, speech intelligibility, spectro-temporal modulation, support vector machine (SVM), Gaussian mixture model (GMM)

1. Introduction

Oral, head and neck cancer refers to the cancer that exists in the upper aerodigestive tract which includes the lips, the oral cavity, the nasal cavity, the pharynx, and the larynx. Those cancers in the brain, the eye, the esophagus, the thyroid gland, as well as many other sites of the head and neck, are not usually classified as head and neck cancer. It represents 3% of all cancers in the United States and is the 6th most common cancer worldwide [1]. Recent studies have shown a five-fold increase of squamous cell carcinoma of the tongue in young men and a six-fold increase among young women [2]. Depending on their tumor size, location and stage, those cancer patients are often treated by radical surgery (such as glossectomy and laryngectomy), radiology, chemotherapy or a combination of above treatments. As a result of those treatments, the anatomical structures and properties of speech organs such as tongue and vocal folds in those patients may change. Therefore, the critical function of speech in those patients might be more or less impacted negatively, along with other functions such as swallowing. Previous studies have shown reduced articulation skills in oral cancer patients which affects both consonants and vowels [3][4]. Voices of laryngectomy patients are often characterized with reduced prosody and rough voice quality [5]. These aspects often lead to reduced speech intelligibility and speech from those patients is often considered disordered or pathologic.

The rehabilitation of cancer patients with disordered speech is of high clinical interest in terms of improving their life quality. Reconstructive surgery followed by speech therapy is used to improve the patients' voice quality and speech intelligibility. Therefore, the evaluation of speech outcome is very important in the rehabilitation procedure. However, the speech intelligibility is usually assessed using subject perceptual rating by a panel of experts or using subject tests in terms of percentage of words or syllables correctly identified [6]. As a result, it is subjective, time-consuming, and also costly to evaluate the outcome of speech therapy. Therefore, there is a need in clinical application for automatic speech intelligibility evaluation on those head and neck cancer patients. It is expected to automatically produce intelligibility scores strongly correlated to subjective perceptual ratings and can be used at least as an objective support when a perceptual evaluation is not available.

There are a few studies on automatic speech intelligibility assessment on pathologic speech (i.e. [7][8][9]). Major efforts have been made to develop automatic speech recognition (ASR) based systems and correlate the phoneme/word accuracy scores in ASR with the ratings in subject perceptual evaluation. The acoustic features used include traditional mel-frequency cepstrum coefficients (MFCC), prosodic, and phonological features. However, despite those efforts, this topic is not well explored in general and still remains challenging.

As an effort towards automatic intelligibility assessment of pathologic speech, this study applied auditory-inspired spectro-temporal modulations as a front-end acoustic feature set on the INTERSPEECH 2012 Speaker Trait Pathology Sub-Challenge. The use of these features was mainly motivated by the importance of spectro-temporal modulations in speech intelligibility and a distortion in those modulations will result in loss of intelligibility [11]. Furthermore, the spectro-temporal modulation-based features have been proven advantageous in several applications such as speech activity detection [12], speech enhancement [13], and speech intelligibility evaluation in noisy environments [14][15]

In the rest of this paper, we first introduce the auditory-inspired spectro-temporal modulations, and describe the speech pathology sub-challenge (its database, evaluation metric, and the baseline feature set), the modulation feature parameter configurations, and the classifiers we used. Then, we present our sub-challenge results using modulation features, along with the results from the baseline system [10]. Finally, a summary along with future work are given.

2. Spectro-temporal modulations

2.1. The computational auditory model

The spectro-temporal modulation features used in this study are inspired by a computational auditory model, which is based on neurophysiology, biophysics and psychoacoustics at various stages of the auditory system. This auditory model captures basic characteristics of signal processing occurring from the early cochlear stage to the central cortical regions. It generates a multidimensional spectro-temporal representation of the sound. Details of this model and its applications can be found in [16][12][13][14][15]. A Matlab toolbox on this model is available online[1].

Basically, the auditory model consists of two processing stages:

- An early auditory stage which mainly mimics the cochlear and mid-brain signal processing and transforms the speech into an auditory spectrogram with a series of cochlear filters, hair cell transduction and lateral inhibition mechanism.
- A central cortical stage which analyzes the auditory spectrogram to get a spectral and temporal modulation profile through a series of modulation-selective filters. The cortical analysis is essentially equivalent to a 2-D affine wavelet transform of the auditory spectrogram.

2.2. The spectro-temporal modulation features

As the output of the auditory model, the spectro-temporal modulation features are a multidimensional array in four dimensions: time, frequency, temporal modulations (called rate), and spectral modulation (called scale). The model output is usually processed with a window size (usually 250 ms, and a shift size varying from 150 ms to 10 ms which depends on applications) and each window is then time-averaged, yielding a tensor with each element representing the overall modulation at a corresponding scale, rate, and frequency. Sufficient number of filters is required in the model to get a full picture of the signal. As a result, the dimensionality of the feature space is very high (for example, 128 frequencies x ±6 rates x 5 scales = 7680) making statistical model training impractical.

Tensor principle component analysis (PCA) was applied to perform dimensionality reduction and a feature dimension as large as 140 was obtained [12]. Further dimensionality reduction can be implemented using a modified linear discriminant analysis (MLDA) in a two class classification problem as in [15].

3. Database and Methodologies

3.1. Database for the pathology sub-challenge

Details regarding the INTERSPEECH 2012 speaker trait pathology sub-challenge and its database can be found in [10]. The “NKI CCRT Speech Corpus” (NCSC) recorded at the Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute [17] was provided to the challenge participants. This corpus contains recordings and perceptual intelligibility evaluations of 55 speakers (10 female, 45 male) who went through concomitant chemo-radiation treatment (CCRT) for inoperable tumors of the head and neck. Based on a threshold in perceptual rating, each utterance in the database was classified into two classes: intelligible (I) or non-intelligible (NI). The whole corpus was partitioned into three sets: the train set, the development set and the test set. Labels (I or NI) of the first two datasets are provided. The challenge is on the prediction of labels in the test set, i.e., each utterance has to be determined as either intelligible (I) or non-intelligible (NI), a two-class problem.

Unweighted average (UA) recall [18] is used as the official evaluation measure in this challenge. In a two class problem, recall is the computed accuracy for each class and UA is calculated as recall(I)/2+recall(NI) /2. Weighted average (WA) is often calculated. But UA is considered as a better performance measure in the case of unbalanced class distributions.

In addition, the 2012 Speaker Trait Challenge baseline feature set was provided for comparison and it contains 6125 features for each utterance [10]. Besides the low level features related to energy, spectra and voicing, a variety of functional applied to low level features were added.

3.2. Feature extraction and parameter configurations

An energy-based voice activity detection (VAD) was used to remove those silence and noisy regions in speech. The final VAD output is the intersection of energy-based VAD and the VAD specified by the ASR transcripts provided along with the challenge database. Then acoustic features were extracted on those regions specified by VAD.

Three variants of modulation features were tested in this study: **a)** the 7680 dimension modulation feature with averaging over time. This feature is to compare with the baseline feature set provided by the Speaker Trait challenge, **b)** the 140-dimension feature through a tensor PCA which is pre-trained using the TIMIT database [12] , and **c)** a feature set (60 dimensions) through further dimensionality reduction by MLDA. The shift size in feature extraction is 100 ms for SVM but 10 ms for GMM in order to get more feature vectors for training.

3.3. Classifiers

The support vector machine (SVM) and the Gaussian mixture model (GMM) are two back-end classifiers applied in this study, depending on feature dimension and training data size. The libSVM [19] and the MIT Lincoln Lab speaker recognition package (for GMM training and scoring) [20] was used respectively. A coarse grid search for optimal parameters on both classifiers was performed. For SVM, both linear kernel and radial basis function (RBF) kernel were tested. For GMM, a mixture number 64 or 128 was selected due to the limited training data.

4. Results

4.1. Averaged spectro-temporal modulations of speech in the NCSC corpus (Intelligible versus Non-intelligible)

In order to gain insight on the difference in modulation pattern between intelligible speech and non-intelligible speech, averaged spectro-temporal modulations in a range of rate and scale (averaged on time and frequency axis) are computed on the whole NCSC corpus for intelligible and non-intelligible speech separately. For checking consistency across datasets, train and devel sets are also computed separately. As shown in Fig. 1, the speech modulations are bounded by the range between 4-8 Hz in rate and < 4 cyc/octave in scale. These four plots look very similar. However, our cross correlation study

[1] The Matlab toolbox is available online: *http://www.isr.umd.edu/Labs/NSL/Software.htm*

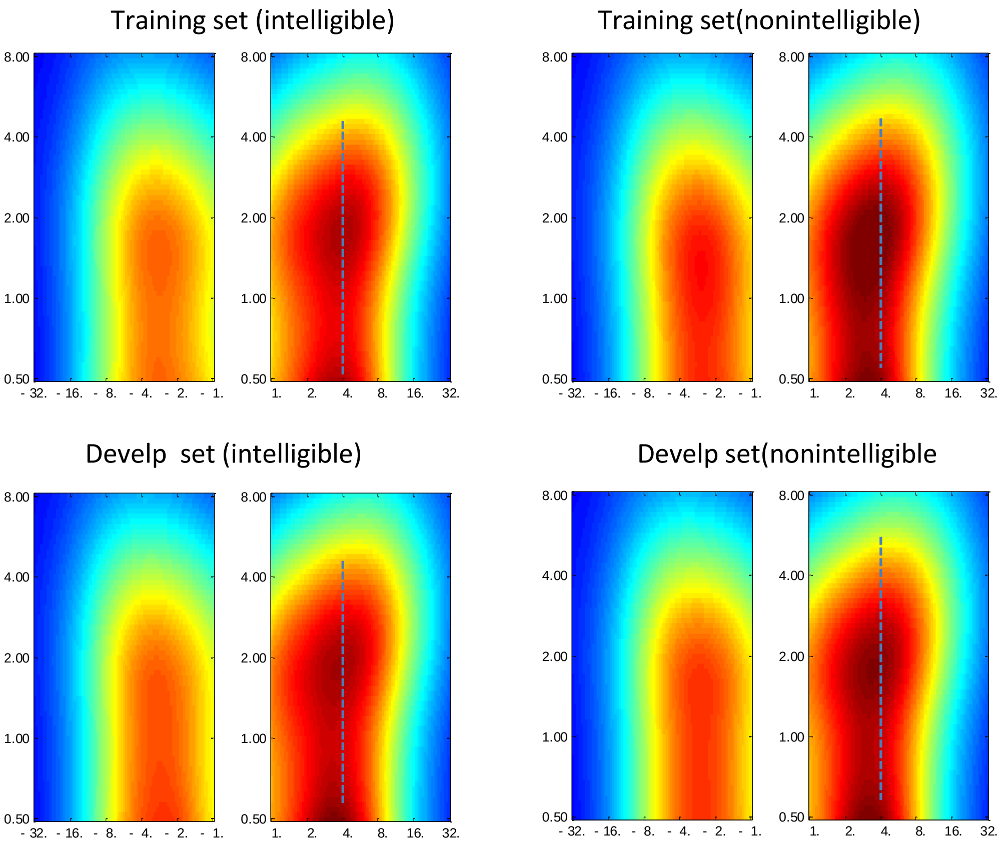


Figure 1. *Averaged spectro-temporal modulations in the NCSC challenge database (In each plot, x-axis is rate in Hz, y-axis is scale in cyc./Octave, color indicates the modulation amplitude, Upper panel: train set, Lower: develop set, Left : intelligible speech, right: non-intelligible speech.*

shows consistently a slightly larger cross correlation coefficient between two plots belonging to the same class than from two different classes. Furthermore, there exists a pattern difference between intelligible speech and non-intelligible speech, that is, while the highest amplitude peaks in intelligible data appears at around 4 Hz (indicated by the dotted line), the peak amplitudes in the non-intelligible data appear at a lower rate (< 4 Hz). This is even clearer when the modulations are further averaged on rate or scale, as shown in Fig. 2. It can be seen that the amplitude peak locations in non-intelligible speech tend to shift to a lower rate and scale (note that the smallest rate in x-axis in Fig. 2a lies in the middle). One explanation for this is that those patients with difficulty in articulation tend to speak slowly and discontinuously. However more future efforts are needed to understand the fundamental difference in spectro-temporal modulations between intelligible speech and non-intelligible speech.

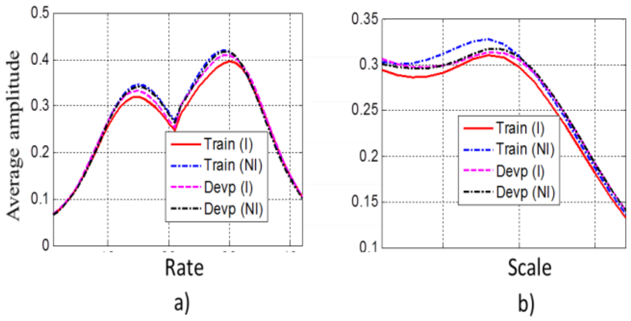


Figure 2. *Averaged modulation on rate* ***a)*** *(from -32 to 32 in Hz) and scale* **b)***(from 0.5 to 8 in Cyc./Octave) in the NCSC challenge database (intelligible or non-intelligible in training or devel set)*

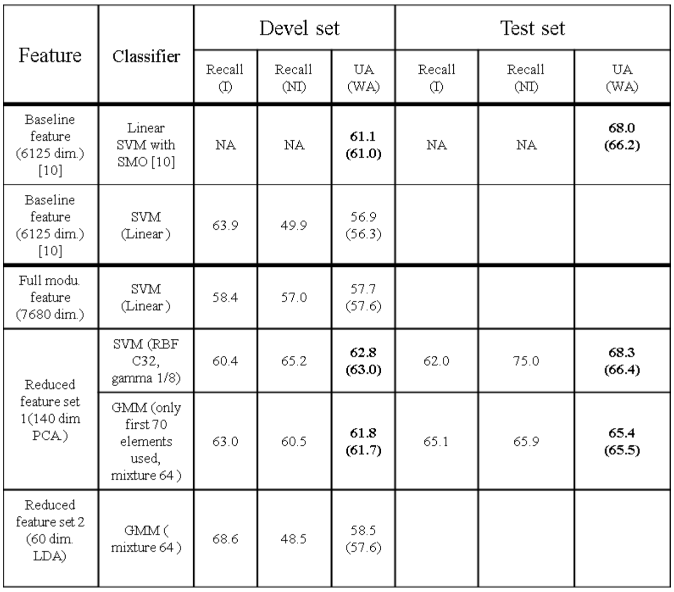
4.2. Experiment results

Our preliminary experiment results on the speaker trait challenge are shown in Table 1 which presents the recalls, the unweighted accuracy, and the weighted accuracy for combinations of feature sets and classifiers. The results on both the development set and the test set are shown in Table 1.

Since the classifiers play big roles in performance, for a fair comparison, SVM with linear kernel was applied on both the baseline feature set and the 7,680 high dimension modulation feature. The RBF kernel was also tested but producing worse results than the linear kernel for the high dimension feature. It can be seen that the 7,680-D modulation feature performs slightly better (57.7 vs. 56.9% UA) than the baseline feature set on the development set. This holds across a range of cost parameter C in SVM. But the difference is small and may not be statistically significant. The 140-D feature set was applied on both SVM and GMM. But only the first 70 elements are used for the latter due to limited amount of training data. In general, GMM is not as good as SVM in this problem, which is mainly due to lacking sufficient training data. The performance of the 140-D feature set using SVM is comparable to the baseline performance on the devel set, 62.8 vs. 61.4% in UW.

The results on the test datasets are shown in Table 1 The performance of 140-D modulation features using SVM is comparable to the baseline performance, 68.3 vs. 68.4% in UW. This shows the effectiveness of spectro-temporal modulations as a feature set in assessment of speech intelligibility.

Table I. The performances of modulation features in the Pathology Sub-Challenge compared to the baseline results by unweighted and weighted average (UA/WA) recall in percent (weighting by number of instances per class)



Feature	Classifier	Devel set			Test set		
		Recall (I)	Recall (NI)	UA (WA)	Recall (I)	Recall (NI)	UA (WA)
Baseline feature (6125 dim.) [10]	Linear SVM with SMO [10]	NA	NA	**61.1 (61.0)**	NA	NA	**68.0 (66.2)**
Baseline feature (6125 dim.) [10]	SVM (Linear)	63.9	49.9	56.9 (56.3)			
Full modu. feature (7680 dim.)	SVM (Linear)	58.4	57.0	57.7 (57.6)			
Reduced feature set 1(140 dim PCA)	SVM (RBF C32, gamma 1/8)	60.4	65.2	**62.8 (63.0)**	62.0	75.0	**68.3 (66.4)**
	GMM (only first 70 elements used, mixture 64)	63.0	60.5	**61.8 (61.7)**	65.1	65.9	**65.4 (65.5)**
Reduced feature set 2 (60 dim. LDA)	GMM (mixture 64)	68.6	48.5	58.5 (57.6)			

The performance of modulation features on the test set can be improved in future on several aspects. First is to optimize the back-end classifier as done in the baseline system [10]. Cross-validation using both train set and development set should get a more generalized classifier on the test set. Second is to fuse the outputs from both modulation features and baseline features based on various classifiers (SVM and GMM) because these two sets of features may provide complementary information. Last is to optimize the modulation feature parameter configurations for a better performance.

5. Summary and future work

This preliminary study applied auditory-inspired spectro-temporal modulations as a front-end acoustic feature set on the INTERSPEECH 2012 Speaker Trait Pathology Sub-Challenge. This was motivated by the premise that a distortion in spectro-temporal modulations of speech will result in loss of intelligibility. The averaged spectro-temporal modulations of the NCSC database were analyzed and it was found that the non-intelligible speech tends to have its modulation amplitude peaks shift towards a smaller rate and scale. Based on SVM and GMM, variants of modulation features were tested on the speaker trait challenge problem and the resulting performances on both the development dataset and the test dataset are comparable to the baseline system. In addition to continuing analysis on the INTERSPEECH 2012 speaker trait challenge in terms of acoustic parameters and classifiers optimization, we would also like to look into fundamental issues on how various phenomena in pathologic speech are represented in spectro-temporal modulations.

6. Acknowledgements

This work was partially supported by a grant from DARPA and NIH grant R01CA133015. All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of NIH and DARPA. The first author would like to thank Sridhar Krishan Nemala for his help on LDA code.

7. References

[1] American Cancer Society, facts and figures, available at http://www.cancer.org/Research/CancerFactsFigures/CancerFactsFigures/cancer-facts-and-figures-2010.

[2] Annertz, K., Anderson, H., Biorklund, A., Moller, T., Kantola, S., Mork, J., Olsen, J. H. and Wennerber, J., "Incidence and survival of squamous cell carcinoma of the tongue in Scandinavia, with special reference to young adults", International Journal of Cancer, vol. 101, pp95-99, 2002.

[3] Sumita, Y. I., Ozawa, S., Mukohyama, H., Ueno, T., Ohyama, T. and Taniguchi, H., "Digital acoustic analysis of five vowels in maxillectomy patients," Journal of Oral Rehabilitation, vol. 29, no. 7, pp. 649–656, 2002

[4] Zhou, X., Stone, M., and Espy-Wilson, C.Y. , "A comparative acoustic study on speech of glossectomy patients and normal subjects", Interspeech 2011, August, Florence, Italy.

[5] Furia , C.L., Kowalski, L. P., Latorre, M.R., Angelis, E. C., Martins, N.M., Barros, A. P. and Ribeiro, K. C., "Speech intelligibility after glossectomy and speech rehabilitation", *Arch. Otolaryngol. Head Neck Surg., v*ol. 127, pp. 877-883, 2001.

[6] Bressmann , T., Sader, R., Whitehill T. L. and Samman, N., "Consonant intelligibility and tongue motility in patients with partial glossectomy", International Journal of Oral and Maxillofacial Surgery, vol. 62, pp. 298-303, 2004.

[7] Bocklet, T., Riedhammer, K., Nöth, E., Eysholdt, U., Haderlein, T. J., "Automatic Intelligibility Assessment of Speakers After Laryngeal Cancer by Means of Acoustic Modeling", J. Voice. 2011 Aug 4.

[8] Maier, A., Haderlein, T., et al., "Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer", EURASIP Journal on Audio, Speech, and Music Processing, Article ID 926951, 2010.

[9] Windrich, M., Maier, A., Kohler, R. et al., "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma," Folia Phoniatrica et Logopaedica, vol. 60, no 3, pp. 151–156, 2008.

[10] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B., "The Interspeech 2012 Speaker Trait Challenge", Proc. Interspeech 2012, Portland, USA, 2012.

[11] Drullman, R., Festen, J., Plomp, R., "Effect of envelope smearing on speech perception", J. Acoust. Soc. Am., 95 (2), 1053–1064, 1994.

[12] Mesgarani, N., Slaney, M., and Shamma, S. A. "Discrimination of speech from nonspeech based on multiscale spectrotemporal modulations", IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, 920-930, 2006.

[13] Mesgarani, N., and Shamma, S. A., "Denoising in the domain of spectrotemporal modulations," EURASIP J. Audio Speech Music Process. 3, 1–8, 2007.

[14] Elhilali M., Chi T, Shamma S. A., "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility", Speech Communication, 41, pp. 331-348, 2007.

[15] Nemala, S. K., Elhilali, M., "A joint Acoustic and Phonological Approach to Speech Intelligibility Assessment", the IEEE International Conference on Acoustics, Speech, and Signal processing, ICASSP'10, pp.4742 - 4745.

[16] Chi T., Ru P.,Shamma S. A., "Multi-resolution spectro-temporal analysis of complex sounds", JASA, 118:887-906, 2005.

[17] Van der Molen, L., Van Rossum, M., Ackerstaff, A., Smeele, L., Rasch, C. and Hilgers, F., "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.

[18] Schuller, B., Batliner, A., Steidl, S., and Seppi, D., "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge", Speech Communication, vol. 53, no. 9/10, pp. 1062–1087, 2011.

[19] Chang, C. C. and Lin, C. J., "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011.

[20] Sturim, D. E., Campbell, W. M., Karam, Zahi N., Reynolds, D., Richardson, F. S. (2009), "The MIT lincoln laboratory 2008 speaker recognition system", INTERSPEECH2009, 2359-62.