Intelligibility Classification of Pathological Speech Using Fusion of Multiple Subsystems

Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab. (SAIL), University of Southern California, Los Angeles, U.S.A.
http://sail.usc.edu

Abstract

Pathological speech usually refers to the condition of speech distortion resulting from atypicalities in voice and/or in the articulatory mechanisms owing to disease, illness or other physical or biological insult to the production system. While automatic evaluation of speech intelligibility and quality could come in handy in these scenarios to assist in diagnosis and treatment design, the many sources and types of variability often make it a very challenging computational processing problem. In this work we design multiple subsystems to address different aspects of pathological speech characteristics. These subsystems are then fused at the binary hard score level (intelligible or not intelligible) using Bayesian networks. Results show that subsystems, such as multiple language phoneme probability system, prosodic and intonational subsystem, and voice quality and pronunciation subsystem, have discriminating power for intelligibility (9.8%, 17.1%, 14.6% higher than by-chance respectively). Noisy-Majority based fusion shows 66.4% accuracy, but the performance improvement by fusion is not made. Also, voice clustering based joint classification is applied to minimize misclassification of the best subsystem, and it shows the best classification accuracy (79.9% on dev set, 76.8% on test set).

Index Terms: pathological speech, intelligibility of speech, fusion of multiple subsystems

1. Introduction

Human vocal production can be disturbed or compromised by a variety of factors, temporarily or permanently, including due to illness and disease, resulting in atypicality in the speech characteristics. This usually results in difficulties in its perception by others, and affects the quality of speech communication, and hence quality of life. For example, head and neck tumors may affect speech quality, and surgical and other interventions can result in other forms of insult to the production system, challenging speech output quality. The factors causing pathological speech quality are many, and so are the resulting characteristics and their impact. This makes their diagnostic assessment, and planning of appropriate treatment interventions, challenging. The state of the art in pathological speech assessment is largely based on subjective judgments of clinical experts. There has been, however, considerable interest over the years to offer objective and automated schemes for measuring and classifying pathological speech quality. This is hoped to offer both improved accuracy and reliability as well as scalability and reduction in the cost of processing.

The perceived intelligibility of pathological speech relies on a number of physical properties of the speech signal. Previous studies have reported a number of features for the automatic assessment of intelligibility for pathological speech [1, 2, 3, 4, 5, 6]. Voice quality features, automatic speech recognition (ASR) based features, perceptual features, phonemic features, prosodic features, and estimated speech production parameters like phonological features have all been reported with their considerable discriminative power in classifying intelligibility of pathological speech.

Despite the large variety, and number, of features developed, the problem continues to be challenged by the wide variability of speech characteristics by disorder. The relationship between the intelligibility in human perception and the atypical variation of signal is still opaque. The wide variability in speaker factors, for example native/non-native, dialectal, gender and age difference, makes this problem even harder. These issues pose a non-trivial challenge in designing automatic system that would work in real world scenarios.

This paper targets the pathology sub-challenge in the speaker trait challenge in Interspeech 2012. The database provided for the challenge includes sentence-level speech in Dutch spoken by patients having head and neck tumors [7]. These patients had gone through concomitant chemo-radiation treatment. Various location and size of tumors may have determined the distortions of their speech, resulting in intelligibility loss of the resultant speech. For example, laryngeal tumors can impede vocal fold movements, causing voice quality distortion [8]. Non-laryngeal tumors in the vocal tract can have negative influence on articulation in speech production, for instance a shift in localization of articulation, modified articulatory tension and compensatory articulation [9]. Inspired by some of these physical properties, we developed and tested a few prosodic features, such as the phoneme-level pitch variation, the duration of voiced segments, and stylization parameters for pitch contour, as well as spectral envelope features encoding articulation details.

ASR-based features are commonly used for intelligibility analysis mainly associated to articulatory malfunction [4, 9, 10]. One approach is to use the output of ASR, like word error rate, for intelligibility assessment, e.g. [3, 4]. Another approach is to use the speech features derived from aligned pair of speech [10]. The drawback with these methods is that they require natural speech data spoken in the same language. Since data collection is generally costly, we suggest an alternative way of using ASR-based feature without any constraint on language. The main idea is to represent acoustic property of each phoneme in pathological speech by using the likelihood score to other phonemes from different languages, acoustic models of which are already readily available.

Lastly we classify multiple samples in group clustered by acoustic similarity. This may help to fix noisy classification decisions. The idea is that if utterances share similar voice characteristics, then they should have a similar pathology annotation.

This paper is organized as following. After a brief explanation about baseline features and systems of this challenge, we will explain each subsystem. Next, the details of joint classification method and fusion scheme will be discussed. Then, we will present the experimental results and discussion of our whole system. Finally, we provide conclusions and directions for future work.

2. Baseline features and systems

This section explains baseline features and systems for the Pathology subchallenge briefly (more details in [7]). This challenge uses "NKI CCRT Speech Corpus," which includes pathological Dutch speech produced by patients who underwent chemo-radiation treatment due to the tumors of the head and neck. 6125 baseline features in total are the functionals of low-level descriptors (LLDs) for each utterance. LLDs consist of energy, spectral and voicing related features, and functionals are various statistics of the LLDs + Δ. The baseline systems are a SVM and a random forest using those baseline features. Their classification accuracy will be provided in the result section.

3. Subsystems

This paper suggests two approaches for intelligible (I) or not-intelligible (NI) classification: (1) the multiple expert subsystem fusion and (2) the knn classifier with joint classification scheme. The detailed description of each subsystem used in the two approaches is provided in the following subsections.

3.1. Multiple language phoneme probability feature

This section explains the probability representations in terms of multiple languages' phonemes and the rationale for its use as pathological speech features. The hypothesis is that the variation of acoustic property of speech signal of one language can be captured in terms of the likelihood of acoustic properties of other languages. The likelihood of "multiple" language phonemes might allow a finer representation, hence helping in discriminating pathological speech sound against naturally well-produced sound.

To test this idea, we adapt the confusion network output generated by the phoneme recognizers in [11] of 3 languages viz. Czech (CZ), Hungarian (HU) and Russian (RU), which are freely available. Each phoneme recognizer generates posterior probabilities which can be used to create a lattice of recognition hypotheses. We convert this, using *lattice-tool*[12], to an approximate confusion lattice or sausage network, for which there exists some ordering. The timestamps are however lost in this process, which we recover by edit distance alignment of the best path from the confusion network with that from the lattice. Epsilon transitions are ignored during this alignment, and a list of time synchronized phoneme hypotheses are obtained in these languages.

Each Dutch phoneme in an utterance is then represented with multiple language phoneme probability as follows. Let $U = \{x^1, x^2, ..., x^N\}$ represent an utterance in train or development set, where x^i is i^{th} phoneme in U; N is the number of Dutch phoneme in U. Then, each x^i in U is represented by a probability vector, $\mathbf{P}_x^i$, consisting of the weighted probability value of each phoneme of CZ, HU and RU: $\mathbf{P}_x^i = [\mathbf{P}_{CZ}^i, \mathbf{P}_{HU}^i, \mathbf{P}_{RU}^i]$, where $\mathbf{P}_{CZ}^i = [w_{CZ_1}^i P_{CZ_1}^i, w_{CZ_2}^i P_{CZ_2}^i, ..., w_{CZ_J}^i P_{CZ_J}^i]$, J: the total number of CZ's phonemes. $w_{CZ_j}^i$ is a weighting term which is the ratio of the duration of overlap between CZ_j, the j^{th} phoneme of CZ and the phoneme boundary of x^1, to the duration of x^1. $\mathbf{P}_{HU}^i$ and $\mathbf{P}_{RU}^i$ are defined similarly. Then, the NI-score for x^i, S_x^i is generated as

$$S_x^i = \begin{cases} \sum_{j=1}^{J'} \frac{P_{CZ_j}^i - \mu_{CZ_j}^i}{\max(\sigma_{CZ_j}^i, \varepsilon)} & \text{if } |P_{CZ_j}^i - \mu_{CZ_j}^i| > (C \times \sigma_{CZ_j}^i) \\ 0 & \text{otherwise} \end{cases}$$

where, $\mu_{CZ_j}^i$ and $\sigma_{CZ_j}^i$ are the mean and the standard deviation of all $P_{CZ_j}^i$ samples in the training data, respectively. Finally, the NI-score for each utterance (S_U) is determined as $S_U = \sum_{i=1}^{N} S_x^i$, which is used as the feature for this subsystem. We test this representation by I/NI classification performance using Linear Discriminant Analysis. Each sentence is tested separately to minimize the variation due to co-articulation in different contexts. Parameters, such as $\mu_{CZ_j}^i$, $\sigma_{CZ_j}^i$, ε and C, are empirically determined for each sentence data based on the classification accuracy on the training set.

3.2. Prosodic and intonational features

We observed that NI speakers often have difficulty in pronouncing a few specific speech sounds, resulting in turn in abnormal prosodic and intonational shape. Additionally, we observed that the pitch trajectory is often not smooth for NI speakers. Motivated by these observations, we design the following phoneme and utterance level features to capture the differences between I and NI classes: the variance of pitch at the phoneme and utterance level, the sum of the pitch L0 norm and utterance duration, pitch stylization parameters obtained by fitting a quadratic polynomial, and nearest neighbor-based confidence estimates for the I/NI classes using mahalanobis distances of MFCCs for each phoneme.

Classification is done per sentence to exclude context-dependent variability of features. Train, dev and test datasets are almost equally divided in terms of sentences, which justifies this approach. While classification within sentence rids us of concerns of the normalization (across sentences) for the features, it also reduces the amount of train data available for each classification tasks. Thus we use a K nearest neighbor (knn) classification approach.

3.3. Voice quality and pronunciation features

Even though voice quality features, such as harmony-noise ratio (HNR), jitter and shimmer, are popularly used for vocal disorder assessment, they have been mostly tested on the prolonged vowel sound, e.g. /AA/. We tested their usefulness in our "read" speech data. From concatenated vowel region signal, HNR sequence is extracted using Praat [19] with default parameters, and jitter local and shimmer sequence are extracted by Opensmile. The utterance level statistics, such as maximum, minimum, [0.1 0.25 0.5 0.75 0.9] quantiles, are estimated.

We also extracted a few pronunciation features, such as cepstral mean normalized 39 MFCCs with 25 millisecond window and 10 millisecond shifting, the 2nd, 3rd and 4th formants in vowel regions, and temporal features, such as average syllable duration, pause duration to the number of syllable ratio, average pause duration, mispronounced phone ratio, and average vowel duration. These features capture the rate of speech, disfluency, and the mispronunciation of phones.

Then, the discriminative power of these features for I/NI classes is tested by knn classification performance on dev set (k=5-15) with brute-force feature selection.

4. Joint classification with speech clustering

In order to group similar speech utterances together, we simply went through a single Gaussian based bottom-up agglomerative hierarchical clustering (AHC) method [20], using LPC based features. The AHC with Kmean post refinement uses generalized likelihood ratio (GLR) as an inter-cluster distance measure [21]. Using the AHC, a final stage of smoothing is applied for correcting some of the predicted labels by knn. This is based on the hypothesis that if speech characteristics of two utterances are very similar, then it is unlikely for an annotator to label them very differently. In other words we assume that the predicted labels should be locally smooth in the voice space. We verify this empirically on the train and dev sets separately by clustering the utterances based on talker voices. Figure 1 shows that labels inside each cluster are usually very similar. Most clusters contain a large percentage of either I or NI labels, except a few near the class boundary. Standard deviation of EWE scores within a cluster are also mostly small. We enforce this constraint by clustering the utterances together on the test set, and then jointly classifying all utterances inside a cluster using a majority voting rule. For example, if 80% of the utterances inside a cluster were predicted to be NI, then labels for all the utterances inside the cluster are changed to NI.

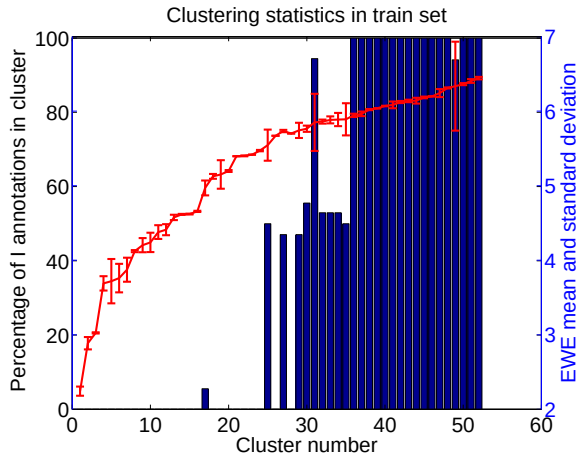

Figure 1: *Distribution of pathology labels (I/NI) and EWE scores in each cluster for the train set. Total 52 clusters are created. The histogram indicates the percentage of I annotations in clusters. The red line plot indicates the EWE mean and standard deviation.*

5. Multiple expert subsystem fusion

We adopt a late score level fusion scheme in this work due to several reasons. First, it allows us to use classification schemes matched to each of the subsystems. This is required because each subsystem is designed to take care of a particular aspect of intelligibility loss and can hence be thought of as a high level descriptor correlated with the binary labels for intelligibility. The contribution of each subsystem to the final accuracy thus, depends on the extent to which they are correlated with the intelligibility labels and also among themselves. Feature level fusion schemes often make oversimplifying assumptions, that might not hold in general.

In this work, we use a Naive Bayes and a Noisy Majority models to fuse the individual subsystem scores. We first learn the parameters for this Bayesian Network viz. the conditional probability tables (CPT) using only the train set. Given these, the inference task is now to calculate the probability $P(\texttt{Int}|S_1, S_2, \ldots, S_N)$ where S_n is the score from the n^{th} subsystem. For training and performing exact inference on this network we use the Bayes Net Toolbox [14].

6. Results and discussions

This section examines the performances of individual subsystems which we suggested in the previous sections. Then, it discusses the benefit of fusion schemes, Naive Bayes and Noisy-Majority fusions. Lastly, it discusses the benefits of the cluster-based joint classification on the result of the best subsystem. Table 1 shows the classification accuracy of each subsystem (trained on train set and tested on dev set).

Table 1: *(Unweighted) classification accuracy of each subsystem trained on train set and tested on dev set. MLPP is the multiple language phoneme probability. 'Pros+Into' is our novel prosodic and intonational features, VQ is voice quality features, Pron is pronunciation features, RF is a random forest on baseline features. Note that SVM and RF are tuned and tested on dev set (provided in the baseline paper [7]). (by-chance: 50%)*

Subsystem	Accuracy (%)
MLPP	59.8%
Pros+Into	67.1%
VQ+Pron	64.6%
SVM (baseline 1)	61.1%
RF (baseline 2)	64.8%

Table 1 shows that our subsystems have useful information for I/NI classification. Average classification accuracy of MLPP subsystem is 59.8% (χ^2 = 3.36, p = 0.07). It indicates that the MLPP subsystem can be useful, but it is not statistically significant for I/NI classification of pathological speech than by chance. The standard deviation of classification accuracy of all sentences is 7.5%, showing that it is context-dependent. The classification accuracy of the sentence-dependent prosodic and intonational subsystem is 67.1% (χ^2 = 22.7, p < 0.01), which is higher than the accuracy of the best baseline system (64.8%) alone. Lastly, the classification accuracy of voice quality and pronunciation features (the best performance was achieved by knn classifier (k=15) with the maximum HNR + all 5 pronunciation features) is 64.6% (χ^2 = 11.40, p<0.01). The significance statistics in parenthesis are obtained by Mc Nemar's chi square test, compared against by chance. These results show that each subsystem is useful for intelligibility assessment, and further the performances of the prosodic and intonational subsystem and the voice quality and pronunciation subsystem are statistically significant.

The scores from some combinations of subsystems are used for the final decision of I/NI by fusion scheme. We tested two Bayesian fusion schemes and joint classification on some subsystems or the best subsystem. Table 2 shows the classification accuracy of them.

Table 2 shows that the best performance is achieved by the joint classification. The joint classification is conducted on the best subsystem, the prosodic and intonational subsystem, with 5 additional features. The 5 additional features are selected from baseline features by brute-force forward feature selection on top

Table 2: *(Unweighted) classification accuracy of final systems (by chance: 50.0%). The best baseline system is a random forest with baseline features in [7]. "dev set" and "test set" indicates the classification accuracies on dev set and test set, respectively. Note that baseline systems are tuned on dev set and tested on dev set or test set, while our systems (Bayesian network fusions and joint classification) are tuned on train set and tested on dev set, or tuned on dev set and tested on test set. The classification accuracy of the best final system is highlighted.*

System		dev set (%)	test set (%)
Baseline SVM		61.1	68.0
Baseline RF		64.8	68.9
Bayesian fusion	Naive Bayes	65.2	
	Noisy-Majority	66.4	
Joint classification		79.9	76.8

of the best subsystem's features. The classification accuracy of the best subsystem + 5 additional features is 71.5%. The joint classification scheme improves classification performance significantly from the knn classification with the prosodic, intonational and 5 additionally selected features (χ^2 = 22.61, p<0.01). It shows that the joint classification reduces the noise from knn's strict classification result. Its classification accuracy is significantly higher than that of the random forest system (χ^2 = 58.26, p<0.01). We also tried to perform the joint classification for a cluster at the posterior level, which yielded 78.8% accuracy. It is less than the hard label fusion results (79.9%). The classification performances of Bayesian fusions on subsystems show even lower accuracy than that of the best subsystem. The reason might be that the data is too small to train the Bayesian networks.

7. Conclusion and future works

This study proposes a few novel features, a novel joint classification scheme for knn classifier, and a Bayesian network based fusion schemes of multiple subsystems for automatic intelligibility assessment. The prosodic and intonational features, multiple language phoneme probability feature, voice quality features and pronunciation features showed discriminating power for binary classification (9.8%, 17.1%, 14.6% higher than by-chance, respectively). Bayesian Network based fusion methods, Naive Bayes and Noisy Majority did not perform well, probably because of insufficient number of data for training Bayesian network. Joint classification based on utterance clustering shows significant improvement of classification accuracy from its subsystem (the prosodic and intonational subsystem) used. It shows that joint classification scheme was able to reduce the number of samples misclassified by knn on the features.

Further analysis is required to study the effect of fusion on each subsystem. Using structure learning on a general Bayesian network system might help in this case. In addition, we would also like to study the effectiveness of other features like inverted glottal pulses or phonological representations that might capture issues in speech production. Lastly, it will be worth to incorporate the knowledge from voice clustering at the training stage, to use similar utterance to train the classifier together instead of treating them as separate samples. This might help to capture the variability within the pathology classes.

8. References

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

[2] Dibazar, A., Berger, T., Narayanan, S., "Pathological Voice Assessment," *In Proc. 28th IEEE EMBS Annual International Conference*, New York, August-September, 2006.

[3] Maier, A. et al., "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," *EURASIP J. Audio Speech Music Process*, 1:1–1:7, 1687-4714, 2010.

[4] Maier, A. et al., "Peaks A System for the automatic evaluation of voice and speech disorders," *Speech Communication*, 51, 2009.

[5] Van Nuffelen, G., Middag, C., De Bodt, M., Martens, J. P., "Speech Technology-Based Assessment of Phoneme Intelligibility in Dysarthria," *International Journal of Language and Communication Disorders*, v44, n5, p716–730, 2009.

[6] Middag, C., Bocklet, T., Martens, J-P., Noth, E., "Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment," *in Proceedings of InterSpeech*, Italy, 2011.

[7] Schuller, B. et al., "The INTERSPEECH 2012 Speaker Trait Challenge," *in Proceedings of Interspeech*, Portland, U.S.A., 2012.

[8] Kazi R. et al., "Electroglottographic comparison of voice outcomes in patients with advanced laryngo-pharyngeal cancer treated by chemoradiotherapy or total laryngectomy," *Int J Radiat Oncol Biol Phys*, 70:344352, 2008.

[9] Jacobi, I. et al, "Voice and speech outcomes of chemoradiation for advanced head and neck cancer: a systematic review," *Eur Arch Otorhinolaryngol*, 267(10): 14951505., October, 2010.

[10] Middag, C. et al., "Dia: a tool for objective intelligibility assessment of pathological speech," *in Proceedings of the 6th International Workshop for Models and Analysis of Vocal Emissions for Biomedical Applications*, 165 - 167, Firenze, 2009.

[11] Schwarz, P., Matejka, P., Cernocky, J., "Hierarchical Structures of Neural Networks for Phoneme Recognition," *in Proceedings of ICASSP*, 325-328, 2006.

[12] Stolcke, A., "SRILM-an extensible language modeling toolkit," *Seventh International Conference on Spoken Language Processing*, 2002.

[13] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, v11, n1, p10-18, 2009.

[14] Murphy, K., "The Bayes net toolbox for matlab," *Computing science and statistics*, v33, n2, p1024-1034, 2001.

[15] Campbell, W.M., Sturim, D.E., Reynolds, D.A., Solomonoff, A., "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *in Proceedings of INTERSPEECH*, 2006.

[16] Li, M., Han, K.J., Narayanan, S., "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, 2012.

[17] Dehak, N. et al., "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol 19, no 4, pp 788–798, 2011.

[18] Chang, C.C., Lin, C.J., "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol 2, no 3, 2011.

[19] Boersma, P., Weenink, D., "Praat: doing phonetics by computer (Version 5.1.08)," Retrieved May 11, 2009, from http://www.praat.org/

[20] Wang, W., Lv, P., Yan, Y. H., "An improved hierarchical speaker clustering," *Acta Acustica*, vol 1, 2008.

[21] Han, K.J., Kim, S., Narayanan, S.S., "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," Audio, Speech, and Language Processing, IEEE Transactions, vol 16, no. 8, pp 1590–1601, 2008.