Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2014) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Automatic intelligibility classification of sentence-level pathological speech☆

Jangwon Kim[a,*], Naveen Kumar[a], Andreas Tsiartas[a], Ming Li[a,1], Shrikanth S. Narayanan[a,b]

[a] *Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, 3710 McClintock Avenue, Los Angeles, CA 90089, USA*[2]

[b] *Department of Electrical Engineering, Computer Science, Linguistics and Psychology, University of Southern California (USC), 3620 McClintock Avenue, Los Angeles, CA 90089, USA*

Received 18 April 2013; received in revised form 31 October 2013; accepted 2 February 2014

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. Although automatic evaluation of speech intelligibility and quality could come in handy in these scenarios to assist experts 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 propose novel sentence-level features to capture abnormal variation in the prosodic, voice quality and pronunciation aspects in pathological speech. In addition, we propose a post-classification posterior smoothing scheme which refines the posterior of a test sample based on the posteriors of other test samples. Finally, we perform feature-level fusions and subsystem decision fusion for arriving at a final intelligibility decision. The performances are tested on two pathological speech datasets, the NKI CCRT Speech Corpus (advanced head and neck cancer) and the TORGO database (cerebral palsy or amyotrophic lateral sclerosis), by evaluating classification accuracy without overlapping subjects' data among training and test partitions. Results show that the feature sets of each of the voice quality subsystem, prosodic subsystem, and pronunciation subsystem, offer significant discriminating power for binary intelligibility classification. We observe that the proposed posterior smoothing in the acoustic space can further reduce classification errors. The smoothed posterior score fusion of subsystems shows the best classification performance (73.5% for unweighted, and 72.8% for weighted, average recalls of the binary classes).

Keywords: Pathological speech; Automatic intelligibility assessment; Dysarthric speech; Head and neck cancer

[2] The SAIL homepage is http://sail.usc.edu.

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +1 213 740 3477; fax: +1 213 740 4651.
E-mail addresses: jangwon@usc.edu (J. Kim), komthnk@usc.edu (N. Kumar), tsiartas@usc.edu (A. Tsiartas), liming46@mail.sysu.edu.cn (M. Li), shri@sipi.usc.edu (S.S. Narayanan).

[1] Now at: The SYSU-CMU Joint Institute of Engineering, Sun Yat-sen University, Guangzhou, China, and the SYSU-CMU Shunde International Joint Research Institute, Guangdong, China.

http://dx.doi.org/10.1016/j.csl.2014.02.001
0885-2308/

1. Introduction

Speech intelligibility assessment is an important measure for evaluation of functional outcomes of surgical and non-surgical treatment, speech therapy and rehabilitation. Patients with vocal disorder or illness that affects the natural control of their vocal organs suffer from compromised communication capability including degraded speech intelligibility. With the overarching goal of improving quality of life, several assessment and treatment approaches, including those targeting speech intelligibility, have been proposed and used in clinical research practice (Kent, 1992). Although the demand for accurate, reliable, and robust intelligibility assessment with low cost is huge (Middag et al., 2009), the state of the art still relies on the perceptual judgment of clinicians and therapists in general. The assessment of pathological speech by professionals can be costly and time consuming, and technology tools can play a supporting role in assisting the experts. In this study we describe an automatic intelligibility assessment system which performs binary intelligibility classification by capturing atypical variation in pathological speech.

The number of vocal disorders and disruption of vocal processes due to illness and disease are many, and therefore there are numerous types of variation possible in pathological speech. For example, differences in the location and size of tumors in the head and neck lead to differences in how the speech signal gets influenced. The tumors of laryngeal cancers affect vocal fold control, resulting in the distortion of voice quality of speech sounds (Kazi et al., 2008). Surgery on lips or nasal cavity can also alter voice quality and resonance or induce hyper-nasality (Hufnagle et al., 1978). Non-laryngeal tumors impede the control of supra-laryngeal organs for speech production (Jacobi et al., 2010). The modified articulatory tension and structural variation, e.g., on the palatal surface and the pharyngeal wall, by non-laryngeal tumors may lead to compromised vocal production, resulting in speech variability.

The signal characteristics of dysarthric speech have also been studied widely in the literature. A previous study showed that change of articulatory manner is associated with dysarthric speech, while variability in articulatory place occurs in both normal and dysarthric speech (Kim, 2010). This study also reported that speakers tend to have more articulatory errors in the production of more complex phones; their findings are based on the complexity metric (for the production of each phone) that Kent (1992) proposed. With regards to the voice source signal, dysarthric speech has been shown to have more variation in utterance-level prosodic features (Kim et al., 2010).

Capturing the wide variability of sources and patterns in pathological speech may require high dimensional acoustic features. These potential variabilities pose challenges for human expert assessment. But, the wide variability in speaker factors, such as gender, age, dialectal, native/non-native difference, makes automated system development even harder. Despite these difficulties, previous studies have reported a range of effective signal cues, including voice quality features (e.g., harmonics to noise ratio, jitter, shimmer), spectral features (e.g., mel-frequency cepstral coefficients, formants), automatic speech recognition scores (i.e. the confidence score of phone or word recognition), perceptual features, phonemic features, prosodic features, and estimated speech production parameters like phonological features (Middag et al., 2009, 2011; Dibazar et al., 2002, 2006; Maier et al., 2010, 2009; Van Nuffelen et al., 2009). Although there have been efforts to demonstrate the effectiveness of some features, e.g., spectral and phonological features, in sentence-level or passage-level data (Maier et al., 2010; Middag et al., 2011; Dibazar et al., 2002), the effectiveness of these signal cues, especially prosodic features and voice quality features, has been studied mostly on datasets collected with simple stimuli, e.g., prolonged vowels and words in limited contexts, presumably in order to reduce the effects of considerable noise in feature extraction due to irregularities of pathological voice.

Although experiments with stimuli of such short duration provide useful segmental information relevant to intelligibility testing with less complexity, data reflecting real-world communication scenarios, e.g., sentence-level or spontaneous speech data, are desirable to ensure both ecological validity and generalizability. The feature characteristics and robustness for single phone- or word-level data might not be consistent with sentence-level speech data due to the high variability and complexity of sentence-level speech production. Hence it is important to examine the effectiveness of conventional pathological speech features (for intelligibility judgment) in the context of sentence-level data and seek suitable novel features which are more effective, robust and reliable for these data.

In addition to feature engineering, this paper also tests several subsystem-fusion schemes for arriving at the final intelligibility decision. Feature-level fusion is one of the most common and easy fusion methods, which combines a variety of features reflecting the possible sources of variability, and often includes feature selection to deal with the curse of dimensionality. This paper examines subsystem-level (decision) fusion, which is another way of handling the high-dimensionality issue. As a by-product, such high-level subsystem fusion offers both an overall intelligibility judgment

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

of a test utterance and quantitative information regarding specific aspects of pathological variability, depending on subsystem design.

The present study also proposes a post-classification smoothing scheme that makes a final decision on a test sample based on the likelihood score of both the test sample itself and other samples in the test set. The main idea is that given the situation that we do not have enough data to cover the wide variability of pathological speech in the train and development sets, we include similarity information in the test set for improving decision making. Also, in real-world scenarios there can be speaker-factor mismatch between the datasets used for model training and parameter tuning, and for testing. It is hence reasonable to make a judgment by including the information underlying in the test set. Additionally our method ensures the consistency of predictions in the acoustic space. This paper will provide details of the proposed posterior smoothing scheme and analyze its behavior as a function of a control parameter for smoothing range.

For experiments we used two different datasets, the NKI CCRT Speech Corpus (van der Molen et al., 2009) and the TORGO database (Rudzicz et al., 2012), for demonstrating the flexibility of our approach across disorders and languages. The NKI CCRT Speech Corpus includes speech audio spoken by native/non-native Dutch speakers (head and neck cancer patients), while the TORGO database includes dysarthric speech in English. Further details of the two datasets are provided in the next section of the paper.

The rest of this paper is organized as follows. In Section 2, we briefly present the NKI CCRT Speech Corpus, the TORGO database and our experimental setup. In Section 3, we describe subsystem design and feature-level fusion, followed by their classification performance on the two datasets. In Section 4, we describe the proposed joint classification and posterior smoothing schemes, followed by the evaluation of their effectiveness for improving intelligibility classification. In Section 5, we present the experimental results of our final system by late score level fusion. Finally, we provide a discussion of the results in Section 6, followed by conclusions and future work directions in Section 7.

2. Databases and experimental setup

2.1. NKI CCRT Speech Corpus

The NKI CCRT Speech Corpus (van der Molen et al., 2009) contains sentence-level speech audio and its perceptual intelligibility score. The speech audio data consist of *read* speech waveforms of 17 Dutch sentences spoken by 55 head and neck cancer patients. The speech audio was collected at three stages based on Chemo-Radiation Treatment (CRT) of patients: before CRT, 10-weeks after CRT and 12-months after CRT. The intelligibility score provided in this corpus is evaluator weighted estimator (EWE) for each utterance, which is computed from the evaluation results of professional listeners. EWE is the weighted mean of evaluation scores from multiple evaluators where the weight is the correlation coefficient of single evaluator's evaluation score to the unweighted mean of evaluations from all evaluators (Grimm and Kroschel, 2005). A total of thirteen native Dutch-speaking speech pathologists participated the evaluation task.

2.2. Pathological speech sub-challenge

The goal of the Interspeech 2012 speaker trait sub-challenge for pathological speech (Schuller et al., 2012) was to build a classification system for binary intelligibility labels on the NKI CCRT Speech Corpus. The binary labels (intelligible (I) and non-intelligible (NI)), were obtained by partitioning the data using the median of the EWE distribution of all the original speech. The sub-challenge participants were required to follow a given data partitioning of train set, development set and test set, each of which was balanced for age, gender and nativeness of the speakers, but not for the number of labels. The challenge provided phone boundaries which were automatically generated by forced-alignment and manual phonetic transcription.

The sub-challenge further offered the performance of two baseline systems and their common feature set. The feature set consists of 6125 utterance-level functionals estimated from prosodic and spectral feature streams and voicing related features, and their derivatives (delta features). The two baseline systems are linear Support Vector Machine (SVM) with sequential minimal optimization and Random Forest. The Interspeech 2012 speaker trait challenge paper (Schuller et al., 2012) offers details of the experimental configuration, feature extraction and baseline systems. Table 1 shows the partitioning of NKI CCRT Speech Corpus for the pathology sub-challenge, which the present study follows mostly,

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

Table 1
Partitioning of the NKI CCRT Speech Corpus.

	Train set	Development set	Test set	Total
Intelligible	384	341	475	1200
Non-intelligible	517	405	263	1185
Total	901	746	738	2385

except that one NI test sample is omitted due to its poor recording quality. Therefore, the number of samples of NI in the test set is 263 in our experiments. Although the present study also presents weighted average recall, all classification performances with the challenge dataset are compared based on unweighted average recall of I and NI classes for consistency with the challenge setup. The initial work presented on this dataset was published by Kim et al. (2012) as an entry, and deemed an eventual winner in the Interspeech 2012 Pathology sub-challenge.

2.3. TORGO dataset

The TORGO database (Rudzicz et al., 2012) contains dysarthric speech audio produced by eight patients (five males and three females) with either cerebral palsy or amyotrophic lateral sclerosis and normal speech audio from seven people (four males and three females) representing the control group. The patients were known to have disruptions in the neuro-motor interface which causes dysarthria. The age range of the patients is from 16 to 50. Although this database was recorded with various types of stimuli, the sentence-level speech audio is used in this study. The prompts used for recording sentence-level speech audio comprise three preselected phoneme-rich sentences sets: the "Grandfather passage," 162 sentences from the sentence intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech, 460 sentences from the MOCHA database, and spontaneously elicited descriptive texts. The details of these prompts and the reason for this selection of prompts are described in the TORGO database paper (Rudzicz et al., 2012). This database provides intelligibility labels in five categorical grades [a, b, c, d, e] which were reduced from an initial 9-point scale, where 'a' is the label corresponding to the best intelligibility and 'e' is the worst.

For all experiments on the TORGO dataset, we used data from 10 speakers (six patients +four people in the control group) which contain phonetic transcripts for each utterance file, because some features in our system need such information. We divided the sentence-level data into binary intelligibility classes (I and NI), following the sub-challenge setup. Table 2 shows the experimental setup, including the number of utterances used in this study. Note that only utterances with phone labels are included for experiments in this study. Since the intelligibility score of sentence-level speech audio of F03 is "a," meaning 'no difficulty,' we assigned label I to her data. We assigned NI label to the data of the other patients, such as F01, M01, M02, M04 and M05. All speakers' data in the control group are considered to be intelligible, so "I" label was assigned to them. For consistency with the experimental setup of the sub-challenge dataset and for reflecting real world scenarios, we trained our systems without the data of test set speakers, using leave-one-subject-out for testing, and used random cross validation for parameter tuning.

The speech audio of the TORGO database was recorded by either a head-mounted microphone or an array microphone. We observed that the speech data, mostly the ones recorded by a head-mounted microphone, often contain considerable channel noise, so we performed spectral noise subtraction, using the VOICEBOX toolbox (Brookes, 2005), before extracting acoustic features. The number of audio samples recorded by an array microphone is 104 (F03

Table 2
Experimental setup with binary intelligibility class for the TORGO database. 'Num. Samples' is the number of samples for each speaker. 'Orig. label' is the original intelligibility assessment of sentence-level speech audio. The original label of the control group (MC01, MC02, MC03 and MC04) is missing, because they are assumed to be intelligible normal speech.

	Intelligible					Non-intelligible				
Speaker	F03	MC01	MC02	MC03	MC04	F01	M01	M02	M04	M05
Orig. label	a					d/e	d/e	d/e	d/e	c
Num. samples	41	186	87	63	157	20	69	50	58	97

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

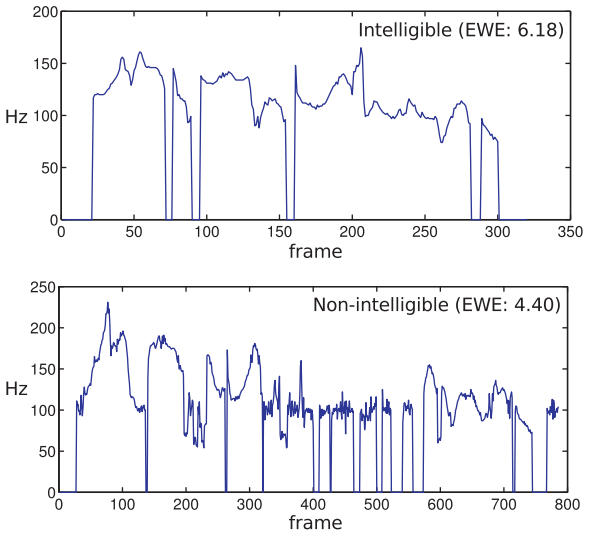

Fig. 1. Example pitch contours of I (top plot is train_046) and NI (bottom plot is train_006) utterances for the same sentence, 'Er leefden eens een koning en een koningin en die hadden maar een kind' in the IS2012 challenge dataset. Both utterances were spoken by male subjects.

and MC03) for I labels and 109 (F01, M01 and M02) for NI labels, while the number of audio samples recorded by a head-mounted microphone is 430 (MC01, MC02 and MC04) for I labels and 155 (M04 and M05) for NI labels. In order to minimize the effects of remaining channel noise on the performance of system evaluation, we performed leave-one-subject-out classification.

3. Subsystem design and testing

3.1. Subsystem for prosody

We observed that NI speakers often have difficulty in pronouncing a few specific speech sounds, resulting in atypical prosodic and intonational shape. Additionally, we observed that the pitch trajectory of the NI speakers' data was often not smooth. Fig. 1 shows the examples of pitch contours of two utterances (one for I and the other for NI) of the same sentence. Motivated by these observations, we designed the following phone- and utterance-level features derived from pitch contours of each utterance.

- Utterance-level features included [0.1 0.25 0.5 0.75 0.9] quantiles, interquartile range of pitch and its delta, normalized L0-norm (the number of non-zero elements divided by the sum of mean duration of each phone in the utterance), normalized utterance duration (utterance duration / the sum of mean duration of each phone in the utterance), the sum of normalized L0-norm ratio and the normalized utterance duration, the z-score of each phone duration, variance of pitch. The sum of mean duration of each phone in the utterance was computed from the entire intelligible speech audio in the train set. For each phone the z-score z_i of sample $x_i \in X = [x_1, x_2, \ldots, x_i, \ldots, x_N]$, where X is all samples of the phone, and N is the number of the samples, is defined as follows: $z_i = (x_i - \overline{X})/S$, where $\overline{X}$ is the sample mean of X and S is the standard deviation of X.
- Phone-level features included the variance of pitch contour and pitch stylization parameters obtained by fitting quadratic polynomials for each phone.

These features were designed in sentence-independent fashion in order to obtain a sufficient number of samples for classifier training. The pitch contour features based on polynomial expansion have also been applied on an age and gender recognition task in a previous study (Li et al., 2013).

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

3.2. Subsystem for voice quality

We tested three types of voice quality features, viz. harmonics to noise ratio (HNR), jitter and shimmer, for intelligibility classification. They have been popularly used for vocal disorder assessment of sustained vowel sound, e.g., /AA/. Since the databases used in the present study are made of sentence-level running speech, we concatenated vowel segments of each utterance instead. Then, we estimated statistics, such as [.05 1.25 5.75 9.95] quantiles, mean, maximum and minimum in the segments for each utterance. We used Praat (Boersma and Weenink, 2009) for extracting HNR with its default parameter set and Opensmile (Eyben et al., 2010) for extracting jitter and shimmer.

3.3. Subsystem for pronunciation

Under the hypothesis that vocal organ malfunction may cause pronunciation variation, thereby contributing to intelligibility loss, we also tested pronunciation features for intelligibility classification. A previous study has shown that formants, cepstral mean normalized 39-dimension Mel-Frequency Cepstral Coefficients (MFCCs), and phone duration are effective in representing pronunciation variation in Dutch (Witt, 1999). Loosely inspired by this study, we developed temporal and spectral feature statistics. The statistics of spectral features include [.05 1.25 5.75 9.95] quantiles, interquartile range, and 3rd order polynomial coefficients (except the residual term) of the first, second and third formants and their bandwidths, and their derivatives for each vowel segment in each utterance. Then, we took the mean of vowel segments (total 132 = 11 statistics × (3 formants +3 bandwidths) × 2). We also estimated the maximum and standard deviation of cepstral mean normalized 39 MFCCs (total 78 = 2 statistics × 39 MFCCs) extracted from utterance-level speech waveforms whose initial and the final silence regions were excluded. The temporal features included average syllable duration, pause duration (without silence before and after speech audio) to the number of syllable ratio, average vowel duration computed with manual phonetic transcription and phone boundaries provided in the database.

3.4. Evaluation of each subsystem and feature-level fusion

In this section, we evaluate the discriminating power of our sentence-level feature sets of individual subsystems and feature-level fusion for intelligibility classification on both NKI CCRT Speech Corpus and TORGO dataset.

Each subsystem described in Sections 3.1–3.3 consists of a large number of features, for the amount of training data available. Hence, the best feature sets were selected based on unweighted average recall of a linear discriminant analysis (LDA) classifier, *k*-nearest neighbor (KNN) classifier and support vector machine (SVM). KNN and SVM have two parameters for tuning, so we followed a standard procedure of joint parameter tuning and forward feature selection. We used Mahalanobis distance metric on the development set, varying *k* from 1 to 20 for the KNN classifier. For the SVM classifier we chose the best kernel function among linear, quadratic, polynomial and Gaussian radial basis functions, in terms of classification accuracy. We used the LIBSVM Matlab toolbox (Chang and Lin, 2001) for SVM model training and testing. For each of the three feature sets (prosody, pronunciation and voice quality), we performed a forward feature selection using the development set, for each value of $I = 1, \ldots, 20$, where I indicates the number of best features. This gave us different selected sets of features corresponding to different values of I. The I of the first locally maximal classification accuracy was chosen and tested on the test set. For the SVM classifier, we performed a forward feature selection with four kernel functions: linear, 2nd-order polynomial, 3rd-order polynomial and radial basis function. For feature-level fusion, feature selection was performed with all features in the three subsystems with each classifier. Table 3 shows the feature selection results of the linear discriminant analysis (LDA) classifier, KNN and SVM, and their classification accuracy on the test set.

Table 3 shows that the best intelligibility classification of an individual subsystem is achieved by the pronunciation feature set with the SVM classifier and 3rd-order polynomial kernel function. The pronunciation features offer the best individual feature performance with the KNN classifier as well, although it fails with the LDA classifier. Feature-level fusion with all features from the three feature sets shows the best performance (69.6% for unweighted average recall, 71.1% for weighted average recall) with SVM with 2nd-order polynomial kernel function. SVM shows better performance than KNN and LDA in general, except in the case of the unweighted average recall of the prosody subsystem on which the best performing classifier is KNN, with similar feature dimensionality. Table 3 suggests that

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

Table 3
Classification results of feature selection and parameter tuning on the pathology challenge data. 'Dim.' is the feature dimension after forward feature selection. 'Acc.' is unweighted (weighted) average recall in % on the test set using I-best features which are chosen based on their classification accuracy on the development set. By-chance is 50% for unweighted average recall and 64.4% for weighted average recall. Note that we evaluated performance based on unweighted average recall to be consistent with the criterion of the IS2012 pathology sub-challenge. 'All' is feature-level fusion with all subsystems' features.

Feature set	LDA		KNN			SVM		
	Dim.	Acc.	k	Dim.	Acc.	Kernel	Dim.	Acc.
Prosody	6	66.0 (69.1)	15	6	66.3 (64.8)	2nd-order poly.	7	64.9 (65.2)
Pronunciation	2	52.9 (50.0)	19	5	66.7 (65.6)	3rd-order poly.	6	67.5 (65.9)
Voice quality	5	62.0 (62.2)	14	6	59.2 (60.7)	2nd-order poly.	5	64.7 (64.8)
All	11	67.9 (70.7)	15	10	66.1 (65.0)	2nd-order poly.	12	69.6 (71.1)

the feature set of each subsystem is useful for intelligibility classification of sentence-level pathological speech from patients with head and neck cancer.

Since the TORGO dataset has an even smaller amount of data than the sub-challenge dataset and since the number of speakers used for experiments was also considerably small, we performed a leave-one-speaker-out cross validation. For each fold, all the data except those from one speaker were used to train classifiers to ensure no speaker overlap between training and testing. Recall that on the TORGO dataset all utterances from a speaker have the same intelligibility rating, which poses overfitting issues while tuning parameters on a development set. This can be understood in terms of bias variance decomposition of classification error. Parameter tuning on the development set seeks to minimize the bias on that set, but in turn causes the model variance to increase, since the amount of data is limited. Hence, on the TORGO dataset, we refer to only results using the LDA classifier, since it does not require any hyperparameter tuning. We still report results using SVM for the sake of completeness. As can be seen, the prosody feature set is less stable with respect to parameter tuning and feature selection on the limited dataset. Table 4 shows the average results of the leave-one-speaker-out cross validation using the optimally-tuned parameters for each fold. For the sake of brevity we omit the optimal parameter values for each fold.

Table 4 shows that the best intelligibility accuracy of an individual feature set is achieved by classifying using only the pronunciation feature set with an LDA classifier. Feature-level fusion with forward feature selection on all three feature sets shows slightly lower performance than the best subsystem, i.e. pronunciation. Classification results in Table 4 support that the feature set of each subsystem is considerably effective for intelligibility classification of sentence-level dysarthric speech.

4. Joint classification of test samples

We attempt to further improve the posterior scores obtained from classifiers by smoothing them on the test set. This is based on the assumption that annotators are less likely to give very different intelligibility ratings to utterances with very similar speech characteristics. In other words, we assume that the predicted labels should be locally smooth in this space of speech features that we describe next.

Table 4
Classification results of feature selection and parameter tuning on the TORGO database. We report both weighted and unweighted average recalls in % because of unequal proportions of the two classes in this dataset. 'All' is feature-level fusion with all subsystems' features. All accuracies in %.

	LDA		SVM	
	Unweighted	Weighted	Unweighted	Weighted
Prosody	82.1	80.3	53.6	51.3
Pronunciation	94.1	93.4	83.8	82.7
Voice quality	71.9	69.0	68.4	73.4
All	93.4	92.9	53.6	51.3

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

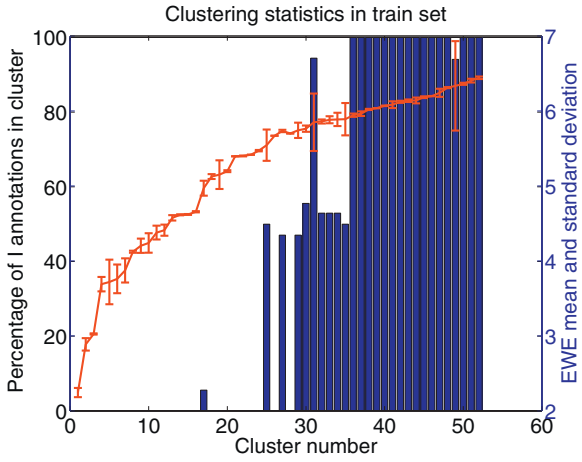

Fig. 2. Distribution of pathology labels (I/NI) and EWE scores in each cluster for the train set. A total of 52 clusters were created. The histogram indicates the percentage of I annotations in clusters. The plot with (red) solid line and error bar indicates the EWE mean and standard deviation. (*Clusters sorted by average EWE score*.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

4.1. Joint classification with clustering

In our previous study (Kim et al., 2012) for the pathology sub-challenge, we verified our assumption separately on the train and development sets by clustering the utterances based on the speech characteristics of subjects. In order to group similar speech utterances together, we performed a single Gaussian based bottom-up agglomerative hierarchical clustering (AHC) proposed by Han et al. (2008) with *K*-means post refinement, using generalized likelihood ratio (GLR) distance. The smoothness constraint was then enforced by an ad-hoc scheme of jointly classifying all utterances inside a cluster using a majority voting rule.

Fig. 2, which is adopted from our sub-challenge paper (Kim et al., 2012), 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 is also mostly small. This figure supports the validity of the assumption.

4.2. Posterior smoothing post classification

In the present paper we propose a more formal smoothing approach which is closer to the general notion of smoothing as a low-pass filtering operation. In other words, we refine the posterior of a test sample as the normalized sum of its neighbors' posteriors weighted by their distances to the test sample in the speech space. The speech space was represented by Line Spectrum Pair (LSP) features (Itakura, 1975). As an alternative linear prediction parametric representation, LSP is closely relevant to the natural resonances or the formants of speech sound, and it is more accurate for the parameterization of the spectral information (Soong and Juang, 1984; Qian et al., 2006). Previous studies have shown the effectiveness of this feature for speech characterization, therefore it is popularly used for the speaker clustering application (Lu and Zhang, 2002; Wang et al., 2008).

We extracted LSP features from each utterance and used GLR distance to perform AHC clustering. The mathematical description of GLR distance is as follows. Suppose that we have a pair of clusters C_x and C_y and that they are represented by two different single Gaussian distributions $N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. They consist of n-dimensional feature vectors with M frames and N frames, where $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M]$ and $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N]$, respectively. If they are from the same speaker, they are merged into one joint cluster C_z with data $\boldsymbol{z} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M, \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N]$ and distribution $N(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$.

GLR distance is based on a hypothesis testing described as follows:

- H_0: $\boldsymbol{x}$ and $\boldsymbol{y}$ follow a joint distribution and are merged together to $\boldsymbol{z}$.

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

- H_1: $\boldsymbol{x}$ and $\boldsymbol{y}$ should follow different distributions and are considered as independent.

Based on these two hypotheses, we can calculate their likelihood ratio, $GLR(x, y)$ as follows.

$$GLR(x, y) = \frac{L(z, N(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z))}{L(x, N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x))L(y, N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y))} \tag{1}$$

Taking into account the single Gaussian distribution, we get the log form of GLR distance (Han et al., 2008).

$$d(x, y) = -\ln(GLR(x, y)) = (M + N) \quad \ln(|\boldsymbol{\Sigma}_z|) - M \quad \ln(|\boldsymbol{\Sigma}_x|) - N \quad \ln(|\boldsymbol{\Sigma}_y|) \tag{2}$$

14-Dimensional LSP features were extracted from the data for every 40-ms Hamming-windowed frame with 20-ms frame shift. The distance between any pair of utterances in the speech space was computed using the aforementioned GLR measure (Han et al., 2008).

Next, we provide a description of the posterior smoothing scheme. If $d(x, y)$ denotes the distance between any two utterances x and y in the speech space, and P_i is the class posterior for the i-th utterance, z_i in the test set, then the smoothed operation is defined as follows:

$$\tilde{P}_i = \frac{\sum_{j=1}^{N} P_j e^{-d(z_i, z_j)/\sigma^2}}{\sum_{j=1}^{N} e^{-d(z_i, z_j)/\sigma^2}} \tag{3}$$

where σ is the bandwidth parameter which controls the scale of smoothing. We chose a Gaussian kernel as our smoothing mask. Note that the extra normalization term in the denominator is necessary in our case, since we only smooth over a finite number of points, instead of a uniform grid of points in the speech space. This ensures a convex combination of the posteriors so that the resulting smoothed posterior is a valid probability. Hence any isolated utterances in the speech space will not be affected by this smoothing operation. Note that the smoothing is only performed over the test set utterances.

To provide the reader with a better intuition as to why the smoothing of posteriors in the test set might be a useful idea, we will try to contrast this method against traditional pattern recognition in which we design classifiers to predict class labels per sample. In other words, if the train set is held constant, the size of the test set does not change the classification results on the test set. This is contrary to what we typically observe. Human experts often perform better when classifying a batch of samples instead of individual samples. This notion of "smoothness" of posteriors on the test set is similar to label smoothing algorithms presented in Zhou et al. (2004) and Zhu and Ghahramani (2002). In other words if two samples are known to be similar beforehand their posteriors are also expected to be similar. In our experiment, we attempt to simulate this idea by imposing smoothness constraints on the test set. By smoothing the class posteriors we try to ensure that a better decision can be made by jointly classifying similar samples.

4.3. Hyper-parameter tuning

We now describe the scheme used for tuning the hyper-parameter σ in Eq. (3), which controls the scale of smoothing of the posteriors. σ is tuned using a binary divide and conquer scheme that tries to find the parameter for which classification accuracy on the development set is maximized. Test set posteriors are then smoothed using this value of σ. Fig. 3 shows an example plot of the classification accuracy as a function of σ. It can be seen from the figure that there is an optimal range of σ for which posterior smoothing improves classification accuracy.

4.4. Results

Table 5 shows the classification accuracy, contrasting results before and after posterior smoothing. It shows that classification accuracy (unweighted average recall) improves as a result of the posterior smoothing in most cases, except the cases of prosody subsystem and feature-level fusion with the LDA classifier and feature-level fusion with the SVM classifier in terms of only unweighted classification accuracy. The gains in the improved cases are significant at the 5% level by the Mc Nemar test for most of the classifiers, except for the voice quality feature set with the SVM.

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

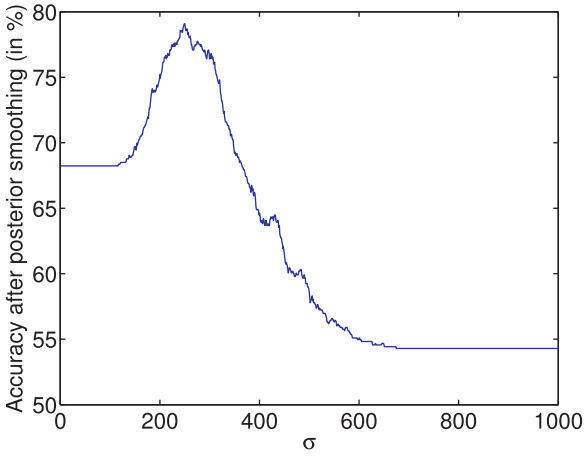

Fig. 3. Sample variation in accuracy with change in hyperparameter σ on the development set for the prosody feature set as an example.

Table 5
Results show the unweighted (weighted) average recall of I and NI labels on test set of sub-challenge dataset before and after posterior smoothing. The hyperparameter σ was learned on the development set. 'All' is feature-level fusion with all subsystem' features. All accuracies in %.

Posterior smoothing	KNN		SVM		LDA	
	Before	After	Before	After	Before	After
Prosody	66.3 (64.8)	69.5 (73.2)	64.9 (65.2)	71.3 (75.5)	66.0 (69.1)	65.3 (69.1)
Pronunciation	66.7 (65.6)	71.0 (71.7)	67.5 (65.9)	71.7 (71.3)	52.9 (50.0)	56.3 (53.0)
Voice quality	59.2 (60.7)	65.9 (68.0)	64.7 (64.8)	66.3 (66.0)	62.0 (62.2)	68.9 (71.7)
All	66.1 (65.0)	72.0 (73.9)	69.6 (71.1)	69.3 (72.3)	67.9 (70.7)	67.2 (70.7)

5. Late score level fusion of multiple subsystems

This section discusses the intelligibility classification performances of our final fusion system. The classification accuracy of the final system on the pathology sub-challenge dataset is compared to the two baseline systems provided in the sub-challenge. We fused each subsystem at the score level using an SVM based fusion scheme. The posteriors obtained from each subsystem are used to train an SVM model to predict the binary intelligibility label on the development set. We use this model for the fusion of class posteriors on the test set.

Table 6 shows the classification accuracy of all final systems. The best performance (73.5% for unweighted and 72.8% for weighted) is achieved by subsystem fusion with smoothed posterior of an SVM classifier. However, subsystem fusion does not always improve classification performance over the best subsystem. For example, the subsystem fusion on KNN's posterior after smoothing shows 0.4% lower unweighted average recall (even though it shows 2.1% higher weighted average recall) of the pronunciation subsystem with KNN's posterior, presumably due to the different bias

Table 6
Unweighted (weighted) average recall in % of final systems (by chance: 50.0% for unweighted, 64.4% for weighted) on the test set of pathological speech challenge database (NKI-CCRT data).

System	Accuracy (%)
Baseline SVM	68.0 (66.2)
Baseline Random Forest	68.9 (67.5)
Subsystem fusion (KNN's posteriors, no smoothing)	70.0 (69.2)
Subsystem fusion (SVM's posteriors, no smoothing)	66.9 (67.8)
Subsystem fusion (smoothed posteriors of KNN)	70.6 (73.8)
Subsystem fusion (smoothed posteriors of SVM)	73.5 (72.8)

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

to I/NI classes between development and test sets. After posterior smoothing, both the feature-level fusion system and subsystem fusion system show better performance than the two baseline systems given in the sub-challenge.

6. Discussion

The performance of the system is evaluated by classification performance on the binary intelligibility labels. However, intelligibility score on at least a five point scale or in percent is more relevant to clinical practice. One possible way is to use a support vector regression (Smola and Schölkopf, 2004) for the final subsystem fusion and generate such final score output. Correlation measure between perceptual intelligibility score and system output can be used for evaluating such final score output (Maier et al., 2009; Middag et al., 2011).

The present study showed the effectiveness of voice quality features for automatic intelligibility assessment in the given datasets. One possible speculation is that atypicality of the speech signal, which is associated with speech intelligibility, due to vocal illness, especially in the larynx, is captured by voice quality features. Although it is well known that laryngeal illness can affect voice quality in speech, there is, however, few works studying the relationship between voice quality and perceptual intelligibility of pathological speech signal. In fact, voice quality has been disregarded in the study of speech intelligibility assessment in general. There is no consensus on the relations between intelligibility and other speech quality factors in the literature. Preminger and Van Tasell (1995) reported that, in the perspective of both multi-dimensional and uni-dimensional views of speech quality, a few perceptual dimensions of speech quality are associated with intelligibility. Particularly, this study suggested that certain speech quality dimensions, such as loudness, listening effort and total impression, are predictable from speech intelligibility score. The present study provides evidence (through experimental results on limited datasets) for the hypothesis that voice quality is also associated with speech intelligibility.

The sentence-dependent features in our earlier system (Kim et al., 2012) are changed to sentence-independent features in the present paper. For example, classification for the subsystem of prosodic and intonational features is done for each sentence in the previously proposed system, while it is done for all sentences in the present paper. The benefit of using sentence-independent features is to secure a larger amount of available training data so that the trained models can have information of the more atypical variability with less concern about over-training. Classification results of individual subsystems with sentence-independent features show their significant effectiveness (shown in Table 3) and they are higher than classification results of sentence-dependent features. For example, in the prosody subsystem, sentence-independent features show higher classification accuracy than sentence-dependent features (66.3% for sentence-independent features, 64.3% for sentence-dependent features) when the KNN classifier is trained in the train set, tuned in the development set, and tested on the test set.

Smoothing of posteriors post classification was suggested to jointly exploit all the test data during classification. This notion of smoothness of posteriors is often exploited in semi-supervised learning (Zhu and Ghahramani, 2002; Zhou et al., 2004) to learn structure from unlabelled data. For convenience, this structure is often encoded in the form of an affinity matrix computed using a kernel function like the Radial Basis Function (RBF) on the features. Since an affinity matrix only depends on pairwise distances, it allows extension to arbitrary distance metrics like GLR (Han et al. (2008)) which might be more intuitive for the problem.

It is worth emphasizing here that in spite of using features related to speaker clustering for computing GLR, the posterior smoothing does not bias the classification task to speaker identity. This is because the challenge data set was designed to prevent any speaker overlap between the train, development or test data (Schuller et al. (2012)). Since the proposed posterior smoothing technique only uses samples from the test set, speaker related characteristics from train data are not transferred to the classification model. Instead, the posterior smoothing technique tries to overcome the shortcomings of the classification system by using information from other samples that are deemed pairwise similar according to the GLR criterion. This method has a direct analogy to low pass filtering by convolution using a filter mask and hence it is called smoothing.

7. Conclusion and future work

This study shows the effectiveness of the automatic intelligibility assessment method we propose, which includes novel sentence-level features and a classifier posterior smoothing scheme. The sentence-level features are designed to capture atypical variation in prosody, pronunciation and voice quality in pathological speech. Pronunciation features

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

were especially found to be promising with a non-linear classifier. Our proposed smoothing scheme was shown to enhance the consistency of class prediction over the test set, resulting in classification accuracy improvement of individual subsystems, feature-level fusion systems, and late score level fusion systems (final systems) in most cases.

Further analysis is required to study the effect of various fusion schemes on each subsystem. This can be accomplished using a Bayesian network system through structure learning on a general Bayesian network system. In addition, it would be interesting to explore the usefulness of other features, e.g., inferred glottal pulses or phonological representations that might capture issues in speech production. It would also be worth investigating data-driven subsystem design for late score level fusion. This might help to capture the variability within the pathology classes.

Acknowledgments

This work was supported by NSF IIS-1116076, NIH DC007124 and DARPA.

References

Boersma, P., Weenink, D., 2009. Praat: doing phonetics by computer (version 5.1.05) [computer program], Retrieved from: http://www.praat.org/ (01.11.11).

Brookes, M., 2005. VOICEBOX: Speech processing toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2 (April (3)), 27:1–27:27.

Dibazar, A., Berger, T., Narayanan, S., 2006. Pathological voice assessment. In: 28th Annual International Conference of Engineering in Medicine and Biology Society. IEEE, 2006, September, pp. 1669–1673.

Dibazar, A., Narayanan, S., Berger, T., 2002. Feature analysis for automatic detection of pathological speech. In: IEEE Proceedings in Joint EMBS/BMES Conference, vol. 1, pp. 182–183.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile – the Munich versatile and fast open-source audio feature extractor. In: Proceedings on ACM Multimedia, pp. 1459–1462.

Grimm, M., Kroschel, K., 2005. Evaluation of natural emotions using self assessment manikins. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 381–385.

Han, K., Kim, S., Narayanan, S., 2008. Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. IEEE Transactions on Audio, Speech, and Language Processing 16 (November (8)), 1590–1601.

Hufnagle, J., Pullon, P., Hufnagle, K., 1978. Speech considerations in oral surgery: Part II. Speech characteristics of patients following surgery for oral malignancies. Oral Surgery, Oral Medicine, Oral Pathology 46 (3), 354–361.

Itakura, F., 1975. Line spectrum representation of linear predictor coefficients of speech signals. Journal of the Acoustical Society of America 57, S35.

Jacobi, I., Molen, L., Huiskens, H., Rossum, M., Hilgers, F., 2010. Voice and speech outcomes of chemoradiation for advanced head and neck cancer: a systematic review. European Archives of Oto-Rhino-Laryngology 267 (October (10)), 1495–1505.

Kazi, R., Venkitaraman, R., Johnson, C., Prasad, V., Clarke, P., Rhys-Evans, P., Nutting, C.M., Harrington, K.J., 2008. Electroglottographic comparison of voice outcomes in patients with advanced laryngopharyngeal cancer treated by chemoradiotherapy or total laryngectomy. International Journal of Radiation Oncology, Biology, Physics 70 (2), 344–352.

Kent, R.D., 1992. Intelligibility in Speech Disorders: Theory, Measurement and Management, vol. 1. John Benjamins Publishing Company, Philadelphia, PA.

Kim, H., 2010. Frequency of consonant articulation errors in dysarthric speech. Clinical Linguistics and Phonetics 24, 759–770.

Kim, H., Hasegawa-Johnson, M., Perlman, A., 2010. Acoustic cues to lexical stress in spastic dysarthria. In: Proceedings of Speech Prosody, pp. 1–4.

Kim, J., Kumar, N., Tsiartas, A., Li, M., Narayanan, S.S., 2012. Intelligibility classification of pathological speech using fusion of multiple subsystems. In: Proceedings of Interspeech. International Speech Communication Association (ISCA), Portland, OR, 2012 September.

Li, M., Han, K.J., Narayanan, S., 2013. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. Computer, Speech and Language 27 (January (1)), 151–167.

Lu, L., Zhang, H.-J., 2002. Real-time unsupervised speaker change detection. In: Proceedings of 16th IEEE International Conference on Pattern Recognition, pp. 358–361.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E., 2009. PEAKS – a system for the automatic evaluation of voice and speech disorders. Speech Communication 51 (May (5)), 425–437.

Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., Schützenberger, A., Schuster, M., 2010. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. EURASIP Journal on Audio, Speech, Music Processing, 1:1–1:7.

Middag, C., Bocklet, T., Martens, J., Nöth, E., 2011. Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In: Proceedings of Interspeech. International Speech Communication Association (ISCA), Florence, Italy, pp. 3005–3008.

Middag, C., Martens, J., Van Nuffelen, G., De Bodt, M.,2009. DIA: a tool for objective intelligibility assessment of pathological speech. In: Models and Analysis of Vocal Emissions for Biomedical Applications, 6th International Workshop. Firenze University Press, pp. 165–167.

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001

Middag, C., Martens, J.-P., Van Nuffelen, G., De Bodt, M., 2009 January. Automated intelligibility assessment of pathological speech using phonological features. EURASIP Journal on Advances in Signal Processing, 3:1–3:9.

Preminger, J.E., Van Tasell, D.J., 1995. Quantifying the relation between speech quality and speech intelligibility. Journal of Speech and Hearing Research 38 (3), 714–725.

Qian, Y., Soong, F., Chen, Y., Chu, M., 2006. An HMM-based Mandarin Chinese text-to-speech system. In: Chinese Spoken Language Processing. Springer, Berlin Heidelberg, pp. 223–232.

Rudzicz, F., Namasivayam, A.K., Wolff, T., 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation 46 (4), 523–541.

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., 2012. The interspeech 2012 speaker trait challenge. In: Proceedings of Interspeech. International Speech Communication Association (ISCA).

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14 (August (3)), 199–222.

Soong, F., Juang, B., 1984. Line spectrum pair (LSP) and speech data compression. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 9, pp. 37–40.

van der Molen, L., van Rossum, M., Ackerstaff, A., Smeele, L., Rasch, C., Hilgers, F., 2009. Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views. BMC Ear. Nose and Throat Disorders 9 (1), 10.

Van Nuffelen, G., Middag, C., De Bodt, M., Martens, J.P., 2009. Speech technology-based assessment of phoneme intelligibility in dysarthria. International Journal of Language and Communication Disorders 44 (5), 716–730.

Wang, W., Lu, P., Yan, Y., 2008. An improved hierarchical speaker clustering. ACTA Acustica 33, 9–14.

Witt, S.M., 1999. Use of speech recognition in computer-assisted language learning. Cambridge University, Cambridge, UK (Ph.D. thesis).

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B., 2004. Learning with local and global consistency. Advances in neural information processing systems 16 (16), 321–328.

Zhu, X., Ghahramani, Z., 2002. Learning from labeled and unlabeled data with label propagation. In: Technical Report CMU-CALD-02-107. Carnegie Mellon University.

Please cite this article in press as: Kim, J., et al., Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.001