Towards an ASR-free objective analysis of pathological speech

Catherine Middag[1], *Yvan Saeys*[2], *Jean-Pierre Martens*[1]

[1]ELIS, Ghent University, Belgium
[2]VIB, Ghent University, Belgium
catherine.middag@elis.ugent.be

Abstract

Nowadays, intelligibility is a popular measure of the severity of the articulatory deficiencies of a pathological speaker. Usually, this measure is obtained by means of a perceptual test, consisting of nonconventional and/or nonconnected words. In previous work, we developed a system incorporating two Automatic Speech Recognizers (ASR) that could fairly accurately estimate phoneme intelligibility (PI). In the present paper, we propose a novel method that aims to assess the running speech intelligibility (RSI) as a more relevant indicator of the communication efficiency of a speaker in a natural setting. The proposed method computes a phonological characterization of the speaker by means of a statistical analysis of frame-level phonological features. Important is that this analysis requires no knowledge of what the speaker was supposed to say. The new characterization is demonstrated to predict PI and to provide valuable information about the nature and severity of the pathology.

Index Terms: objective intelligibility assessment, pathological speech, phonological features, running speech

1. Introduction

As communication has been acknowledged as an essential part of life, also for persons with disordered speech, speech intelligibility diagnosis and monitoring in the course of therapy have become increasingly important in the past decade. Where speech intelligibility is traditionally measured in perceptual tests with professional listeners (speech therapists), recent work has demonstrated that an ASR can take over the role of the human listener and enable the design of an automatic and objective assessment. In previous work [1, 2], we showed that it is possible in this way to automate the Dutch Intelligibility Assessment (DIA) [3, 4], a test in which the listener must identify for each monosyllabic word utterance the missing phoneme in a word template. The DIA is shown to yield a reliable intelligibility at the phoneme level, and the automated DIA offers an objective score which correlates well with that perceptual score.

A first problem with the test is that phoneme intelligibility (PI) is only correlating moderately with the ability to communicate in a more realistic situation where running speech is the speech mode [4, 5]. A second problem is that especially children tend to make reading errors because they often misread a nonsense target word as a more common existing word. These errors obviously induce a negative bias in the speaker's intelligibility. Because of these problems, we envisage an automated test that utilizes running speech and that is robust against reading errors, hesitations, etc. of the speaker.

We contemplate that it would be difficult to use an ASR in such a test because it would encounter large difficulties to handle the Out-Of-Vocabulary (OOV) words that are induced by reading errors. This is already the case for reading errors made by normally speaking children [6], let alone for errors that are made by children with a speech disorder. This is why we have searched for a novel ASR-free approach.

A first attempt to predict speech intelligibility without an ASR was made by Bocklet et al. [7]. In that attempt, a speaker verification approach was adopted: a GMM was trained for every speaker, and the parameters of that GMM were used as features from which to predict the speaker's intelligibility. The approach proposed here relies on phonological feature detectors that were trained once and for all on a sufficiently large corpus of normal speech. They offer a phonological feature representation that is presumed to relate to articulatory dimensions, and that is therefore potentially interesting for a more detailed assessment of the speaker's articulation problems.

2. Speech corpus

In order to train and evaluate the envisaged models, we conduct experiments with a part of the Dutch Corpus of Pathological and Normal Speech (COPAS). The corpus was constructed in the project Speech Algorithms for Clinical and Educational applications (SPACE) [8]. It contains recordings of 318 Flemish speakers, pathological as well as control speakers. For a majority of the speakers, only the DIA was recorded, but for 122 speakers, we also have recordings of a read text passage. The recorded passage is a Dutch equivalent of passages like "Grandfather" (for English) or "Nordwind und Sonne" (for German). It contains the Dutch standard text of "Papa en Marloes", consisting of 8 phonetically rich sentences.

Of the 122 speakers 6 have a voice disorder, 26 have a hearing impairment, 48 have dysarthria, 15 have laryngectomy, 1 has glossectomy and 26 are normal (control) speakers. Perceptual PI scores (derived from the DIA recordings) are available for all speakers, but no running speech intelligibility (RSI) scores. More details on the recording conditions and the severities of the speech disorders can be found in [9].

3. Objective analysis

The system proposed here comprises four processing stages which are depicted in Figure 1. The incoming speech $s(n)$ (an utterance of the complete text) is first subjected to a short-term acoustic analysis. The output vectors X_t of that analysis are converted into phonological feature confidence vectors Y_t. Each phonological feature confidence is subjected to a statistical analysis with the aim to derive a compact description of the feature pattern over the whole utterance. The descriptions of the different features are merged into a vector Z which characterizes the speaker. This vector is finally supplied to an Intelligibility Prediction Model (IPM) [1] that predicts the RSI, a

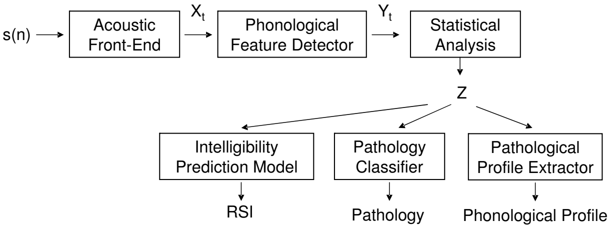

Figure 1: *Schematic diagram of speech production.*

pathology classifier which determines the type and severity of the pathology, or a pathology profile extractor which yields a profile of the speaker in one or more low-dimensional articulatory subspaces (see result section).

3.1. Acoustic analysis in the Front-End

The front-end analysis is a standard MFCC-analysis (Mel-Frequency-Cepstral Coefficients) with a frame size of 30ms, and a frame shift (hop size) of 10ms. Per frame t it provides a vector X_t consisting of 13 features: 12 MFCC coefficients and a log-energy. To minimize the influence of the microphone, Cepstral Mean Subtraction is performed.

3.2. Phonological feature extraction

The vectors X_{t-1}, X_t and X_{t+1} are supplied to a phonological feature extractor whose outputs refer to 14 distinct phonological features describing voicing, place of articulation, turbulence, nasality, etc. We only extract phonological features that can emerge from local information only. This means that modulation features like ”trill” are currently not considered yet.

The phonological feature extractor is composed of Artificial Neural Networks (ANNs) which have been trained on a corpus of read speech by 174 normal speakers (GoGeN, [10]). The corpus is supplied with a phonetic segmentation and labeling. To prepare the training data, we first create a table containing the canonical values of the 14 phonological features of each phone. Eleven phonological features, like nasality for instance, are of a ternary nature: they can either be 1 (feature is on/present), 0 (feature is irrelevant) or -1 (feature is off/absent). Continuously valued features, like the vowel property “front-back”, are also modeled as ternary features with the zero being used for all values differing from the extremes. Three features (voicing, silence and turbulence) are of a binary nature (only having +1 and -1 as acceptable values).

Each ternary feature is represented by two outputs which are derived by a cascade of two single-output ANNs: the first ANN discriminates between irrelevant (0) and relevant (-1 or +1), the second one between absent (-1) and present (+1). We experienced that this ANN-tandem yields a more accurate distinction between present and absent than a single ANN.

Given that there are 11 ternary and 3 binary features, the output Y_t consists of 25 continuously valued components, each representing the degree of confidence for the presence/absence and the relevance/irrelevance of one phonological feature at frame t.

3.3. Phonological characterization of the speaker

A statistical analysis of each component of Y_t is performed in order to construct a phonological feature vector per speaker. We hypothesize that the fluctuations in a phonological feature pattern (over time) can reveal an articulatory deficiency of the speaker, in spite of the fact that the phonetic nature of the frames is unknown (that knowledge would have to come from an ASR). Obviously, this may not be true anymore if the utterance is too short to have a phonetic content that is sufficiently representative of speech in general.

If a component of Y_t either describes a binary feature or the relevancy of a ternary feature, the statistical analysis runs over all frames. If it is the presence/absence of a ternary feature, it is only analyzed over the frames with a positive relevancy. We derive both frame-level and segment-level statistics. To that end we define relevant (irrelevant) segments as intervals of more than 2 consecutive frames where ternary feature is relevant (irrelevant). Similarly, we define positive (negative) segments as intervals where a relevant feature is present (absent). For every component of Y_t, the following features are derived:

1. mean value,
2. standard deviation,
3. percentage of relevant/positive frames,
4. percentage of relevant/positive segments,
5. mean over all relevant/positive frames,
6. mean over all irrelevant/negative frames,
7. mean duration of a relevant/positive segment,
8. mean duration of an irrelevant/negative segment,
9. mean of the maximum in a relevant/positive segment,
10. mean of the minimum in an irrelevant/negative segment,
11. mean time needed to reach the maximum within a relevant/positive segment,
12. the mean time needed to reach the minimum within an irrelevant/negative segment.

Most features aim to reveal whether the speaker has difficulties in realizing clear presence/absence/irrelevance distinctions, but others are more looking for problems related to the switch between presence and absence. In total a speaker is characterized by 25 x 12 = 300 features, and one can expect high correlations between some of them.

4. Experimental study

To test the potential of the above ASR-free phonological feature generation, we have built and evaluated an IPM for predicting PI, as well as a classifier for detecting whether there is a pathology or not.

Since the training of more detailed classification models would require more reliable targets than we have available right now, we do not present finer classification results. Instead, we have conducted a qualitative analysis of the speaker features to demonstrate that they effectively encode the type and severity of the speaker's pathology.

4.1. Training and validation procedure

For the training and validation of our models we adopt a five-fold cross validation scheme, and the listed results are averages (over the five folds) of the root mean square errors (RSME) between computed and target outputs. To investigate the potential of our newly developed feature set, we have conducted experiments on 122 speakers. We examine two feature sets: the new ASR-free phonological features which have been derived from the running speech recordings of the speakers, and the best ASR-based features [2] that have been derived from the DIA monosyllabic word recordings of these speakers. The ASR-based models are now developed on 122 speakers whereas in [2] they were developed on a larger corpus of 211 speakers.

Table 1: *RMSE between the computed and the target PI scores for the old ASR-based and the new ASR-free method.*

Model	ASR-based	ASR-free
Linear regression	8.9	9.8
Support Vector regression	8.8	9.7

The ASR-based features (128 in total) consist of context-dependent phonological features, derived with a phonologically-based ASR, and context-independent phonemic features, derived with a state-of-the-art HMM-based ASR. We formerly showed [2] that an IPM based on these features can attain an accurate prediction of PI.

4.2. Prediction of the RSI

We first like to demonstrate that a reliable RSI prediction on the basis of the new speaker features is possible. However, since COPAS just provides PI scores we can only gather indirect evidence. Relying on the known correlation between PI and RSI [11] we contemplate that if our speaker features can be converted to PI, they can be converted to RSI as well, provided that perceptual RSI scores are available for model training. We develop two models: one based on ensemble linear regression with feature selection and one based on Support Vector Regression (SVR).

For the training of an ensemble linear regression model we create ten random divisions of the training fold: one part for regression coefficient estimation and an equally large part for model assessment. As a result, we get ten models per training fold. These models are then recombined into one single model which is finally evaluated on the validation fold. This process is embedded in an iterative scheme that, starting from the best subset of 3 features, utilizes the individual model assessments (on part of the training fold) to identify which is the best feature to add to the feature subset that was chosen in the previous iteration.

The SVR is achieved by a Support Vector Machine (SVM) with a gaussian kernel. During the training of the SVR on a particular training-validation partition, we select the learning parameters (kernel parameters, fault threshold) by means of a grid search based on internal cross validations on five folds defined within the training part.

The results of the different PI prediction models can be found in Table 1. They confirm that the novel method can predict PI intelligibility in a reliable way. The ASR-free and the ASR-based models compete rather well, especially since the ASR-based method is actually favored because its features are extracted from the DIA recordings which gave rise to the target PIs. Another finding is that there are no significant performance differences between the two statistical learners. Maybe these differences will pop up when larger datasets become available.

4.3. Detecting speech disorders

Our second objective was to predict from the speaker features whether the speaker has a speech disorder or not. For this classification we investigate three modeling strategies: logistic regression, SVM and Ripper [12], a simple rule induction system. For every SVM training, the learning parameters are selected according to the method developed by [13].

The results attained by the different models are depicted in Table 2. Apparently, the features derived from the ASR-based system outperform the new features for all model types. It is not yet clear right now why this is the case. Although for the ASR-free features, the Ripper method cannot compete with SVM, we hope that the very simple and compact rules derived by Ripper will help us to gain more insight in the main indicators of pathology. Take for instance the following rule, extracted for the case of ASR-free features:

Table 2: *Error percentage for pathological versus normal speaker classification based on ASR-based and ASR-free features.*

Model	ASR-based	ASR-free
Logistic regression	17.2	22.1
SVM	8.2	13.0
Ripper	10.6	22.9

IF (A16 $\geq$ 0.28) and (A32 $\geq$ 0.99)
THEN Class=Normal, ELSE Class=Pathologic.

With this rule, 58% of the normal and 99% of the pathological speakers are correctly classified. The rule emphasizes the importance of the relevance related features: A16 (standard deviation of relevancy for alveolar) and A32 (maximum relevancy for nasality). Roughly speaking, the rule points out that pathological speakers have problems to realize sufficiently positive evidence for the alveolar feature, and that they often sound hypernasalized.

We have also tried to go one step further and to discriminate between the different pathologies. While a discrimination between individual pathologies seems to be difficult (error rates of the order of 35%, probably due to an under representation of some classes), the discrimination between a specific pathology and the control group seems to be feasible. For the two 'large' groups (dysarthria and hearing impairment), both the ASR-based or ASR-free features lead to perfect discrimination using an SVM based classifier.

4.4. Speaker profile extraction

An argument in favor of our approach is that it works with features that are closely related to articulatory dimensions, and that a limited number of features might be sufficient to get a more detailed characterization of the type and severity of the articulatory problems of a certain speaker. In order to get evidence in support of this argument, we have examined all 2-dimensional subspaces of the ASR-free speaker feature space. We utilize Linear Discriminative Analysis (LDA) to learn the distinction between normal speakers and either hearing impaired speakers or laryngectomees. We identify interesting subspaces as subspaces in which this distinction can be made with high accuracy. We observe that for the ASR-free features, classification accuracies of up to 92% can be achieved, whereas with ASR-based features the accuracy is limited to 87%.

Figure 2 shows a scatter plot of the hearing impaired and the normal speakers in the subspace of "mean of alveolar" and "mean time to reach the maximum within a relevant segment for nasality". The figure confirms the findings of [14, 15] that hearing impaired speakers sound hypernasal. The depicted feature combination is the best in four of the five folds, and the second best in the fifth fold.

For the laryngectomees, features for voicing appear to be very discriminating, as could be anticipated. However, less expected, all the best feature pairs also contain at least one feature concerning turbulence, referring to fricative and plosive sounds.

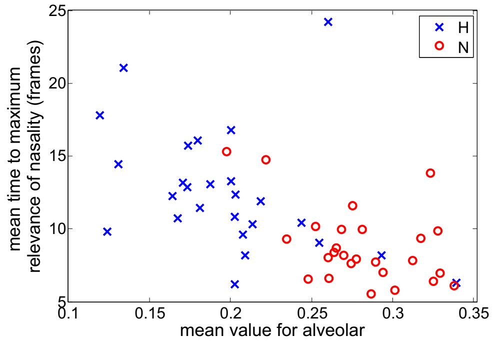

Figure 2: *Scatter plot of control speakers (N) and hearing impaired speakers (H) in the most discriminative subspace of the speaker feature space.*

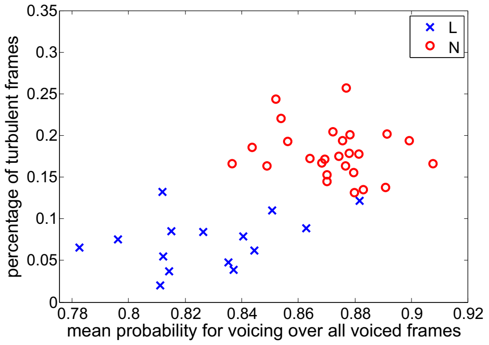

Figure 3: *Scatter plot of control speakers (N) and laryngectomees (L) in the most discriminative subspace of the speaker feature space.*

This complies with the fact that in the ASR-based approach, we found fricative to be an important feature. Although this needs further investigation, Figure 3 seems to support the hypothesis that this may have something to do with the inability of laryngectomees to switch between voiced and unvoiced sounds. This would mean that they have difficulties realizing turbulence during a relatively short period in the vicinity of voiced sounds.

5. Conclusions and future work

We have proposed a novel ASR-free methodology for the objective assessment of pathological speech. The method is based on phonological features that are closely related to articulatory dimensions, and it analyses running speech as a natural speech mode. Important is that it does not require a transcription of the target speech. This way, the method is anticipated to be resistent to reading errors made by the speaker. Unfortunately, we have insufficient running speech recordings at our disposal, and no corresponding running speech intelligibility scores for these recordings. Therefore, we have only been able to conduct an exploratory investigation of the potential usability of our method as the corner stone of a future automated assessment of speech pathology on the basis of running speech.

At present, our system already achieves a very good prediction of phoneme intelligibility as a proxy for running speech intelligibility. Furthermore, it can clearly distinguish a specific type of pathology from normal speech in two-dimensional subspaces that can be identified automatically.

The advantage of our features in terms of knowledge discovery is also exemplified, thereby proving that the ASR-free phonological features effectively point to specific articulatory dimensions that might explain the specific pathology.

Future work will focus on the further development of a robust diagnosing system that offers an intelligibility prediction as well as a speaker profile. From such a profile one could then retrieve objective information about the progress of a certain patient in the course of a therapy. As to enable us to build more sophisticated intelligibility prediction models we will collect more running speech recordings and collect running speech intelligibilities for these recordings. We will further improve our speaker feature extraction and also include segmental features (e.g. “trill”) and supra-segmental features (e.g. intonation patterns) in the analysis.

6. References

[1] C. Middag, G. Van Nuffelen, J. P. Martens, and M. De Bodt, “Objective intelligibility assessment of pathological speakers,” in *Proceedings of the International Conference on Spoken Language Processing, Brisbane, Australia*, 2008, pp. 1745–1748.

[2] C. Middag, J. P. Martens, G. V. Nuffelen, , and M. D. Bodt, “Automated intelligibility assessment of pathological speech using phonological features,” *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 9, 2009.

[3] M. De Bodt, C. Guns, and G. V. Nuffelen, *NSVO: Nederlandstalig Spraakverstaanbaarheidsonderzoek*. Herentals: Vlaamse Vereniging voor Logopedisten, 2006.

[4] G. Van Nuffelen, M. De Bodt, F. Wuyts, and P. Van de Heyning, “Reliability and clinical relevance of a segmental analysis based on an intelligibility assessment,” *Folia Phoniatrica et Logopaedica*, vol. 60, pp. 264–268, 2008.

[5] R. Kent, G. Weismer, J. Kent, and J. Rosenbek, “Toward phonetic intelligibility testing in dysarthria,” *Journal of Speech and Hearing Disorders*, vol. 54, pp. 482–499, 1989.

[6] J. Duchateau, K. Demuynck, and H. V. Hamme, “Evaluation of phone lattice based speech decoding,” in *Proceedings of the European Conference on Speech Communication and Technology, Brighton, U.K.*, 2009, pp. 1179–1182.

[7] T. Bocklet, T. Haderlein, F. Hönig, F. Rosanowski, and E. Nöth, “Evaluation and Assessment of Speech Intelligibility on Pathologic Voices based upon Acoustic Speaker Models,” in *Proceedings of the 3rd Advanced Voice Function Assessment International Workshop*, 2009, pp. 89–92.

[8] http://www.esat.kuleuven.be/psi/spraak/projects/SPACE/.

[9] G. Van Nuffelen, C. Middag, M. De Bodt, and J. P. Martens, “Speech technology based assessment of phoneme intelligibility in dysarthria,” *International Journal of Language and Communication Disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[10] K. Demuynck, D. V. Compernolle, C. V. Hove, and J. P. Martens, *Een Corpus gesproken Nederlands voor spraaktechnologisch Onderzoek. Final Report of CoGeN Project*. ELIS UGent, Gent, 1997.

[11] G. V. Nuffelen, “Speech intelligibility in dysarthria,” Ph.D. dissertation, Antwerp University, 2009.

[12] W. W. Cohen, “Fast effective rule induction,” in *ICML*, 1995, pp. 115–123.

[13] M. Varewyck and J. P. Martens, “Data-driven model selection for support vector classifiers with a gaussian kernel,” *IEEE Transactions on Systems, Man, and Cybernetics: Part B*, to appear.

[14] S. Leder and J. Spitzer, “A perceptual evaluation of the speech of adventitiously deaf adult males,” *Ear and hearing*, vol. 11, no. 3, pp. 169–75, 1990.

[15] S. G. Fletcher and D. A. Daly, “Nasalance in utterances of hearing-impaired speakers,” *Journal of Communication Disorders*, vol. 9, no. 1, pp. 63 – 73, 1976.