Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment

Catherine Middag[1], *Tobias Bocklet*[2], *Jean-Pierre Martens*[1], *Elmar Nöth*[2]

[1]Department of Electronics and Information Systems, Ghent University, Belgium
[2]Chair of Pattern Recognition, University of Erlangen-Nuremberg, Germany
catherine.middag@elis.ugent.be tobias.bocklet@informatik.uni-erlangen.de

Abstract

Intelligibility is widely used to measure the severity of articulatory problems in pathological speech. Recently, a number of automatic intelligibility assessment tools have been developed. Most of them use automatic speech recognizers (ASR) to compare the patient's utterance with the target text. These methods are bound to one language and tend to be less accurate when speakers hesitate or make reading errors. To circumvent these problems, two different ASR-free methods were developed over the last few years, only making use of the acoustic or phonological properties of the utterance. In this paper, we demonstrate that these ASR-free techniques are also able to predict intelligibility in other languages. Moreover, they show to be complementary, resulting in even better intelligibility predictions when both methods are combined.

Index Terms: pathological speech, objective intelligibility assessment

1. Introduction

Speech therapy is an increasingly important discipline in our society. It plays a major role in improving communication skills and helping people in speech rehabilitation. In particular for patients with a speech pathology, speech therapy can help them to regain (some of) their vocal and articulatory abilities. To measure these improvements, speech intelligibility is a widely used measure. Apart from the commonly used perceptual intelligibility tests, some automatic and thus intrinsically more objective intelligibility methods have been developed, making use of ASR systems, such as the PEAKS platform [1] and the DIA tool [2]. Those methods proved to constitute a reliable and objective alternative for the existing perceptual tests. They can now be consulted on-line by speech therapists and act as a kind of objective (unbiased) listener who never gets familiar with the patient's speech. In both PEAKS and DIA, the speaker's utterance is compared with the prompted text. In PEAKS, this comparison is made by an ASR with a small test-dependent dictionary, limited to the words in the prompted paragraph. The basic idea is that the ASR system has increasing trouble recognizing pathologic speech with an increasing degree of pathology. Intelligibility is measured as the percentage of correctly recognized words. In DIA, the comparison is obtained by aligning the speech to the list of prompted words, and by deriving from that alignment a set of speaker features which are on their turn transformed into an objective intelligibility score. As much as the above methods have proven to work well for the task they were designed for, they encounter problems when the pathological speaker starts to make hesitations and reading errors, as it often happens with children speakers. Clearly, these errors should have no impact on the intelligibility, but they do introduce out of vocabulary words which cause an alignment or a recognition system to derail. To circumvent this lexical problem, a new philosophy of deriving speaker features was conceived. In this philosophy, no ASR and especially no lexicon is employed.

A first attempt to predict speech intelligibility without an ASR was made by Bocklet et al. [3]. In that attempt, a speaker verification approach is adopted: a GMM is trained for every speaker, and the parameters of that GMM constitute a supervector from which to predict the speaker's intelligibility. This method led to high correlations between computed and perceptual intelligibility scores for a German dataset consisting of 85 partially laryngectomized speakers. As only acoustical properties of the speech are used, this approach is claimed to be language-independent.

Another ASR-free approach was presented in [4]. It relies on a statistical analysis of the feature patterns emerging from phonological feature detectors which are trained on normal speech. This method attained promising results on a Flemish dataset consisting of 122 speakers with a variety of pathologies. The phonological feature set is claimed to be independent of the used language. Moreover, it is presumed to relate directly to the articulatory dimensions of speech, and as such, it may be suitable for conducting a more detailed assessment of the speaker's articulation problems in a later stage.

As both ASR-free approaches capture different characteristics of the speech signal, it makes sense to investigate whether combining them is beneficial. In this paper, we investigate whether the feature sets are really language-independent and whether combining them leads to a model that outperforms the individual models. For this purpose, we conduct experiments on the two datasets that were formerly used to test the individual approaches that were presented in [3] and [4] respectively.

2. Datasets

In this section we describe the two datasets we used for training and evaluation of the individual and combined models.

2.1. German Partial Laryngectomees (GPL)

The dataset used in [3] contains recordings of 85 patients who suffered from cancer in different regions of the larynx. 65 patients had already undergone partial laryngectomy and were recorded on average 2.4 months after surgery, while the remaining 20 were still awaiting surgery. Each person read the German version of "The Northwind and the Sun". This is a phonetically balanced text composed of 108 words (71 disjunctive) and containing all phonemes of the German language. The text is frequently used in speech therapy [5] in German speaking coun-

tries. More details about the recording conditions can be found in [3].

Five phoneticians and speech scientists rated every speaker's intelligibility according to a 5-point Likert scale [6]. The average of these five ratings is used as a reference during the automated intelligibility assessment.

2.2. Flemish Pathological Speech (FPS)

This dataset is a part of the Dutch Corpus of Pathological and Normal Speech (COPAS), made publically available through the Dutch Language Union[1]. It contains recordings of 318 Flemish speakers, pathological as well as non-pathological control speakers. For a majority of the speakers, only recordings of the isolated word test (DIA) are available, but for 122 speakers there are also recordings of the standard Dutch text passage "Papa en Marloes" [7] consisting of 8 phonetically rich sentences. We have performed our experiments on these recordings. More details on the recording conditions and the severities of the speech disorders can be found in [8].

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 phoneme intelligibility (PI) scores (derived from the DIA recordings) are available for all speakers, but there are no perceptual running speech intelligibility (RSI) scores. In this paper, we consider the PI score as a proxy for the RSI score.

3. Feature extraction

The acoustic front-end computes the standard Mel Frequency Cepstrum Coefficients (MFCCs) which are very popular in the field of automatic speech recognition [9]. The frame rate is 10 ms and the frame size is 25 ms. For each frame t, the first 12 MFCCs and the log energy are retained, together with their first and second order derivatives, to constitute a 39-dimensional feature vector X_t. To minimize the influence of the microphone, Cepstral Mean Subtraction (CMS) is applied for each sentence or word list.

Based on these features, the approaches in [3] and [4] are applied to create two ASR-free speaker feature sets: an acoustical and a phonological feature set.

3.1. Acoustical ASR-free features (AC-ASRF)

The first system, described in [3], is based on a statistical modeling of the acoustic space of the speaker and on the assumption that the acoustics of pathologic speakers differ from those of non-pathological speakers. The degree of pathology is measured as the distance between the pathologic speaker model and a reference speaker model. The speaker model is a Gaussian Mixture Model (GMM) representing all available MFCC vectors $\boldsymbol{X}$ of the speaker. The reference model is a speaker-independent GMM that is trained on speech of healthy speakers. This model is usually referred to as the *Universal Background Model* (UBM).

The UBM is trained in an unsupervised iterative manner by the *Expectation-Maximization* (EM) algorithm [10] in 5 iteration steps. It computes likelihoods by means of

$$p(X|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \omega_i p_i(X|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}). \quad (1)$$

[1] http://www.inl.nl/en/producten

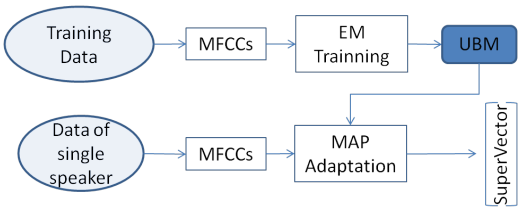

Figure 1: *Composition of the GMM-based supervector by concatenation of the mean vectors*

where the ω_i, $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma_i}$ denote the weights, the mean vectors and the covariance matrices of the different mixtures. The number of Gaussian densities M is set to 128.

A speaker model is derived by adapting the parameters of the UBM to the data of the speaker. Since only a limited amount of data is available for each diagnosed speaker, only the mean vectors $\boldsymbol{\mu_i}$ are adapted. This is accomplished by means of *Maximum A Posteriori* (MAP) adaptation [10]. The adapted means constitute a so-called GMM-based supervector by a simple concatenation of them (see Figure 1). This vector is expected to represent well the acoustic space of the speaker. It is referred to as AC-ASRF and it is composed of 39 x 128 = 4992 individual features.

3.2. Phonological ASR-free features (PH-ASRF)

While the first system models the speaker in the acoustic space, the second system builds a set of phonological features that representation the speech of a speaker.

For every frame t, the acoustic feature vectors X_{t-1}, X_t and X_{t+1} are supplied to Artificial Neural Networks (ANNs) which have been trained on a corpus of read speech by 174 normal (non-pathological) speakers (GoGeN, [11]). The network outputs represent 14 frame-level phonological features describing voicing, place of articulation, turbulence, nasality, etc. on a local time scale. Until now, we only extract features that can emerge from local information alone, which excludes e.g. features like "trill" which only emerge on a longer time scale.

Eleven frame-level phonological features (e.g. nasality) are of a ternary nature.

In most cases this means that they can either be 1 (feature is on/present), 0 (feature is irrelevant) or -1 (feature is off/absent). In some cases, like for the continuously valued "front-back" property of a vowel, the value 1 refers to "front", -1 to "back" and 0 to anything else between those 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 emerging from two cascaded single-output ANNs: the first one discriminates between 0 and ± 1, the second one between -1 and +1. All ANN outputs computed for frame t are collected in a 25-dimensional vector Y_t .

For each speaker, a statistical analysis of the temporal evolution of the individual components of Y_t is performed to construct a 300-dimensional phonological feature vector PH-ASRF describing that speaker. The temporal analysis calculates characteristics such as the mean, standard deviation, percentage of positive, zero and negative values, maximum and minimum values, mean time needed to reach a maximum or minimum, etc.

The idea is that temporal fluctuations in the components of Y_t can reveal articulatory deficiencies of the speaker, regardless of the phonetic nature of the frames (information that would normally be provided by an ASR). Obviously, this will only be true is the utterance spoken by the speaker has a sufficient phonetic richness, and is sufficiently representa-

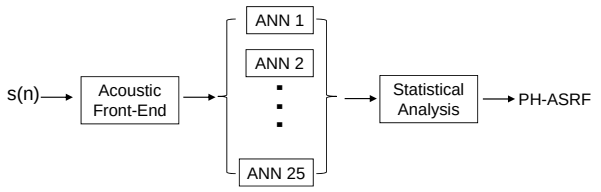

Figure 2: *Schematic diagram of the phonological feature extraction process.*

tive of speech in general. Some PH-ASRF features may reveal whether the speaker has difficulties in realizing clear presence/absence/irrelevance distinctions, whereas others are more looking for problems related to the switch between presence and absence. A flow chart of the phonological feature extraction can be found in Figure 2.

4. Experimental setup

Starting from the two ASR-free speaker feature sets, four different intelligibility prediction models (IPMs) per dataset were created, as depicted in Figure 3. Two of them (IPM 1 and IPM 2) consider only one of the feature sets and constitute the baseline model. The two others (IPM 3 and IPM 4) employ combinations of the both feature sets.

4.1. Training and validation procedure

For the training and validation of our models we adopted a leave-one-out cross validation scheme. We tried two statistical learners for every IPM: one based on ensemble linear regression (ELR) with feature selection [4] and one based on Support Vector Regression (SVR) [12].

For the training of an ensemble linear regression model we created 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. The mean output of these ten models is then evaluated on the validation fold. This process is embedded in an iterative scheme that, starting from the best feature, utilizes the individual model assessments to identify which is the best feature to add to the feature subset that was chosen in the previous iteration.

The SVR experiments were conducted in Weka [13]. The learning parameters were set to the default values. Gaussian, linear and polynomial kernels of degree 3 were tested.

4.2. Combination of feature sets

There are several ways of constructing an IPM that combines two feature sets. We adopted 2 strategies: early fusion and late fusion. Both are displayed in Figure 3. Early fusion (IPM 3 in the figure) combines both feature sets into one set which is then used for the training of the IPM. Late fusion (IPM 4 in the figure) uses the outputs of two IPMs trained on the individual feature sets and combines the results of these two models to obtain the final result. The combination of both IPMs can again be accomplished in several ways, and can be as complex as performing an extra SVR to map both individual model outputs to the intelligibility score. However, this would require an extra cross-validation loop. Therefore, for this paper, we simply calculate the final intelligibility score as the mean of the individual intelligibility scores.

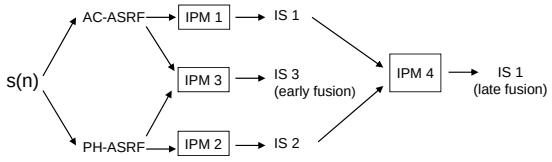

Figure 3: *Early and late fusion. Predictions of the Intelligibility Scores (IS)*

5. Results and discussion

In this section we present the results for the four IPMs in combination with SVR and ELR as the training algorithm. We computed the Pearson Correlation Coefficient (PCC) and the Root Mean Squared Error (RMSE) between the computed and the target outputs as our evaluation measures. The RMSE is expressed in percent of the full scale: 5 for GPL and 100 for FPS. The Wilcoxon signed-rank test [14] is applied to investigate whether differences between results are significant at a confidence level of 0.05.

Table 1 provides an overview of the results obtained for the FPS-dataset. Note that the target outputs for this dataset are actually phoneme intelligibility (PI) scores derived from listening to isolated monosyllabic word utterances. Consequently, there is a certain degree of mismatch between these PI scores and the envisaged RSI scores that would have emerged from listening to the paragraph passage.

Table 1: *PCCs and RMSEs (see text) for the two datasets. In case of SVR, linear kernels are denoted by lin, gaussian kernels by RBF and polynomial kernels by poly followed by their degree. Per dataset, the underlined results denote the reference system, those in bold indicate performances differing significantly from that reference.*

		SVR			ELR	
data	feature set	kernel	PCC	RMSE	PCC	RMSE
FPS	AC-ASRF	lin	70	9.4	**44**	**11.6**
	PH-ASRF	poly3	69	9.5	65	9.8
	early fusion	lin	71	9.2	65	9.8
	late fusion	-	**74**	**8.7**	64	10.2
GPL	AC-ASRF	lin	81	11.0	72	13.0
	PH-ASRF	RBF	81	11.0	69	12.8
	early fusion	lin	81	11.0	73	12.8
	late fusion	-	84	10.4	73	12.6

A first major finding is that SVR clearly outperforms ELR as a learning method. We come back to this later. Looking at the SVR models, it appears that both feature sets, AC-ASRF and PH-ASRF, perform equally well on both datasets. This can be considered as proof of the fact that these two feature sets can be used in a language independent scenario, as claimed but not verified in the original papers where they were introduced. Another result is that early fusion is not capable of exploiting the complementarity of the two feature sets, whereas late fusion can. Late fusion causes a statistically significant improvement on the FPS dataset. This is exemplified by a drop of the RMSE by about 8% relative. However, the improvement on the GPL set is only significant at a confidence level of 0.08. Optimizing the parameters of the SVR training instead of using the Weka default values and adopting a more efficient late fusion technique might further improve the results and lead to significant

difference with a lower p-value on both datasets. That early fusion is not capable of causing any improvement may well be a consequence of the fact that the combined feature set is very unbalanced, namely 4992 AC-ASRF features against only 300 PH-ASRF features.

Note that all models perform better on the GPL than on the FPS dataset. This is owed to the fact that the GPL dataset only comprises laryngectomees. The dominant cause of the diminished intelligibility this type of speakers resides in the diminished amount of voicing that is produced. This type of deviation is obviously easier to model than a more complex articulatory deficiency involving e.g. a combination of problems related to both the manner and the place of articulation. Such complex deficiencies are bound to occur frequently in the FPS-dataset. Another factor might be that the reference scores in the FPS dataset were not measured on the examined utterances, but emerged from separate utterances of another type (isolated words instead of continuous speech).

A striking result, already mentioned casually, is that the AC-ASRF features perform very badly on the FPS dataset when used in combination with ELR. The most likely explanation of this phenomenon is that the AC-ASRF feature set consists of many strongly correlated components, and that the simple strategy of adding one feature at the time is not a valid strategy in that case. To give an example, if the mean vector of a mixture component in the speaker model differs from the corresponding mean vector of the UBM, it is probably important to measure in which direction the mean vector has moved. This direction information is encoded in a linear combination of mean vector components and is not necessarily well reflected in any of the individual components of that vector. Consequently, the feature addition method may fail to add any of these components to the subspace in which the regression will take place. In SVR, the features are always examined together. That the phenomenon is so much more apparent in the FPS dataset than it is in the GPL dataset is probably a consequence of the larger complexity of the envisaged modeling task in the FPS dataset. After all, this set represents multiple pathologies which involve more complex articulatory deficiencies than the GPL set which only contains speech of laryngectomees. The reduced speech intelligibility for these speakers is to a large extent caused by their lack of ability to realize a voiced/unvoiced distinction.

6. Conclusions and future work

Previous work described two different approaches to compute the intelligibility of a pathological speaker without the need for an automatic speech recognizer. Generally speaking, the two methods both follow a tandem approach consisting of an acoustic front-end to extract the traditional MFCC features, a speaker feature generator to create a model of these features that summarizes the articulatory phenomena observed in the speech of that speaker, and an intelligibility prediction model to convert these speaker vectors into an intelligibility score.

Both methods were formerly shown to predict speech intelligibility rather well on the kind of data they were trained on – German and Flemish speech respectively. In this paper, we first of all demonstrate that the features emerging from the two methods compete well with one-another on the two datasets. This also implies that the two speaker feature sets are indeed language independent, as claimed but not verified in the original papers. Secondly, we have shown that combining the two feature sets in one system is beneficial, provided late fusion is employed as the fusion technique.

Since late fusion is achieved by just averaging the outputs of two intelligibility production models, there is still room for improvement. Future work can be directed towards the training of another regression model that can better map the two outputs onto the desired intelligibility score.

As both methods have been proven to work in a language-independent scenario, we can start to explore more datasets covering more languages, in the hope that the combined method will prove to be applicable in general. Finally, it would be interesting to establish whether the method can also be used in a text-independent scenario.

7. References

[1] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS - A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, pp. 425–437, 2009.

[2] C. Middag, J. P. Martens, G. V. Nuffelen, , and M. D. Bodt, "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*, 2009, p. 4.

[3] 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.

[4] C. Middag, Y. Saeys, and J. P. Martens, "Towards an asr-free objective analysis of pathological speech," in *Proceedings of the International Conference on Spoken Language Processing, Tokio, Japan*, 2010, pp. 294–297.

[5] I. P. A. (IPA), *Handbook of the International Phonetic Association*. Cambridge University Press, 1999.

[6] R. Likert, "A technique for the measurement of attitudes." *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.

[7] J. Van De Weijer and I. Slis, "Nasaliteitsmeting met de nasometer," *Logopedie en Foniatrie*, vol. 63, pp. 97–101, 1991.

[8] 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.

[9] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[10] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[11] 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.

[12] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[13] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with java implementations," in *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, 1999, pp. 192–196.

[14] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. CRC Press, 2004.