Objective intelligibility assessment of pathological speakers

Catherine Middag[1], *Gwen Van Nuffelen*[2], *Jean-Pierre Martens*[1], *Marc De Bodt*[2]

[1]ELIS, Ghent University,Belgium
[2]Antwerp University Hospital, Belgium
catherine.middag@elis.ugent.be, gwen.van.nuffelen@uza.be

Abstract

Intelligibility is a primary measure for the assessment of pathological speech. Traditionally, it is measured using a perceptual test, which is by definition subjective in nature. Consequently, there is a great interest in reliable, automatic and therefore objective methods. This paper presents such a method that incorporates an automatic speech recognizer (ASR) for producing features that characterize the pronunciations of a speaker and an intelligibility prediction model (IPM) for converting these features into an intelligibility score. High correlations (about 0.90) between objective and perceptual scores are obtained with a system comprising two different speech recognizers: one with traditional acoustic models relating acoustical observations to triphone states and one using phonological features as an intermediate layer between the acoustical observations and the phonetic states.

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

1. Introduction

Intelligibility is defined as the accuracy with which a listener is able to decode the acoustic signal of a speaker [1]. It is a highly relevant measure for the assessment of pathological speech. Traditionally, clinicians evaluate a patient's intelligibility by means of a perceptual test like [2, 3]. Such a test is tedious and partly subjective. This calls for the development of automatic assessment methods. Recent work [4, 5] has already demonstrated high correlations between the word accuracy of an automatic speech recognizer (ASR) trained on normal speech and a subjective impression of intelligibility (a mean opinion of several human judges).

In this paper, a novel method based on phonemic and phonological feature scores, derived from a forced alignment of the speech with the target text, is presented. It has permitted us to create several automatic intelligibility prediction systems and to measure the agreement between the objective intelligibility scores they produce and the perceptual scores that were available in a pathological speech corpus.

If phonological feature scores can predict intelligibility, what will be proven here, they can most probably also be a basis for more detailed predictions. This would give the advantage that the amount of intelligibility loss due to specific articulatory phenomena like e.g. the horizontal tongue position, the lip movements, etc. can be predicted. This articulatory information would be of immediate relevance to the clinician who wants to design an appropriate therapy and monitor its effectiveness.

2. Subjective evaluation and database

The subjective assessment against which we will compare our objective methods is the Dutch Intelligibility Assessment (DIA) [2], which was constructed to measure intelligibility at the phoneme level. Every speaker reads 50 consonant-vowel-consonant words which are divided in 3 subtests for testing initial consonants (19 words), final consonants (15 words) and medial vowels and diphthongs of Dutch (16 words) respectively. To avoid that the listener (a clinician) gets too familiar with the test items, there are 25 variants of each subtest and each variant contains existing words as well as pronounceable pseudowords. For each test item, the listener must fill in the missing phoneme in a word frame such as ".it" (in case the initial consonant is the target phoneme). Indicating an omission is also allowed. The perceptual intelligibility score is calculated as the percentage of correctly identified phonemes. Previous research [2, 6] showed that the intelligibility scores derived from the DIA are highly reliable (an intraclass correlation of 0.93 and an interclass correlation of 0.91 [6]).

We have collected a database of 10550 consonant-vowel-consonant word recordings (50 words x 211 speakers) produced by 51 control speakers, 60 dysarthric speakers, 12 children with cleft, 42 persons with pathological speech secondary to hearing impairment, 37 laryngectomized speakers, 7 persons diagnosed with dysphonia and 2 persons with a glossectomy. The subjective phoneme intelligibilities of the pathological speakers vary between 28 and 100 percent with a mean of 78.7 percent, while those of the control speakers range from 84 to 100 percent with a mean of 93.3 percent.

3. Objective intelligibility assessment

Opposed to the approach advocated in [4, 5], we propose a two-stage system with an ASR producing a set of speaker features and an intelligibility prediction model (IPM) transforming these features into an intelligibility score. The advantage of this approach is that the ASR can be trained on normal speech whereas the IPM can be trained on pathological speech. We tested two ASRs and different sets of speaker features. Both ASRs use Mel-Frequency-Cepstral Coefficients (MFCC) [7] as inputs (frame size = 30ms, hop size = 10ms, 13 features per frame).

3.1. Speech recognizers

The first system (ASR-ESAT) we consider is the main-stream state-of-the-art ASR [8] developed at ESAT. It is a Semi-Continuous HMM system comprising a large set of state-independent Gaussians and acoustic triphone models. A global phonetic decision tree defines a large number of tied states.

The second system (ASR-ELIS) is developed in our own

department. In a first stage, a neural network based phonological feature extractor [9, 10] extracts 24 binary phonological features concerning voicing, vowel height, manner of articulation, place of articulation etc. per frame. These features are then supplied to an ASR engine with context-independent phone models (some phonemes appear as multiple-phone units). This is acceptable because coarticulations can be handled by the phonological feature extractor which is 'seeing' long (110 ms) time intervals.

3.2. Feature extraction

Since we had access to the global scores of the recognizers as well as to the frame-level scores of the acoustic models they contain, we had the opportunity to let the ASRs produce different types of speaker features, which we briefly describe here.

3.2.1. Word Accuracy (WA)

The simplest speaker feature one can derive from the ASR is the word accuracy (WA), defined as the percentage of correctly recognized words. A word is considered correctly recognized if the target word obtains the highest score. By computing a WA for the full test and for the three subtests A, B and C one obtains a set of four WA-features per ASR. This kind of features is also used in [4].

3.2.2. Log-Likelihood Ratio (LLR)

As the WA is based on a binary decision (word correctly or not correctly recognized), it might be useful to try out a continuous measure to circumvent the effects of discretization. This is done by using the LLR measure. This measure is defined as the log likelihood of the target (correct) word minus that of the best other word. Here too, we can retrieve four LLR-features per ASR.

3.2.3. Phonemic Features (PMF)

It is possible to obtain a richer speaker characterization by analyzing the phonetic segmentation made by the ASR based on the target word. Note that since only the target word is considered, the ASR is actually used in a forced alignment mode. There is evidence [11] that measures derived from such an alignment tend to correlate with intelligibility.

If the aligner assigns frame t with acoustic representation X_t to an acoustic model state s_t, the proposed method first computes the posterior probability $P(s_t|X_t)$ for that frame. In ASR-ESAT, this requires the conversion of likelihoods $f(X_t|s_t)$ to posteriors according to

$$P(s_t|X_t) = \frac{p(X_t|s_t)P(s_t)}{p(X_t)} \quad (1)$$

$$p(X_t) = \sum_{s\in S} p(X|s_t)P(s_t) \quad (2)$$

with S being the set of eligible states and $P(s_t)$ the prior probability of visiting state s_t. In ASR-ELIS, posterior probabilities $P(k, A|X_t)$ of phonological feature k being equal to A (0 or 1) [9] are converted to $P(s_t|X_t)$ according to

$$P(s_t|X_t) = \left[\prod_{k,A_{ck}(s_t)=1} P(k, 1|X_t)\right]^{\frac{1}{N_1}} \quad (3)$$

with $A_{ck}(s_t)$ representing the canonical value of phonological feature k of state s_t and N_1 the number of canonical values that are 1 for this state. Note that ASR-ELIS uses single-state models.

A phonemic feature PMF(f) for phoneme/phone (depending on the ASR) f can then be derived by taking the mean over the posterior probabilities $P(s_t|X_t)$ of all frames X_t assigned to any state s_t belonging to the phoneme or phone f.

Repeating this process for every phoneme/phone gives rise to 40 PMFs for ASR-ESAT and 55 PMFs for ASR-ELIS.

3.2.4. Phonological Features (PLF)

Using ASR-ELIS, it is also possible to compute a set of phonological features PLF(k, A), expressing how well the acoustic observations predict that phonological feature k has a value A (0 or 1) in frames where it is supposed to be equal to A. Those PLFs are derived from the posterior probabilities $P(k, A|X_t)$ of phonological feature k being equal to A in 3 steps:

1. Consider all frames X_t assigned to the state modeling phone f and compute the mean $P(k, A|X_t)$ as a posterior for k having a value A over all occurrences of f, denoted as $P(k, A|f)$.
2. Repeat this for all phones f.
3. Now calculate the PLF(k,A) as the mean of the posteriors $P(k, A|f)$ over all phones f whose $A_{ck} = A$.

Since different speakers have spoken different material (subtest variants) and since our PLF features are based on all phonemes appearing in that material (and not only on the phoneme tested by the perceptual DIA), a simple averaging of scores over frames would have implied that the impact of a phone (e.g. /i/) on the value of a phonological feature (e.g. front) would be variable. By averaging the posteriors $P(k, A|f)$ per phone first, we give equal weights to every phone contributing to PLF(k,A).

In case of ASR-ESAT, one cannot compute a PLF with the same interpretation but one can nevertheless introduce the notion of phonological features by adapting the procedure that delivered the PMFs. This is done in 3 steps:

1. Split the plosives into a part containing the closure and one containing the burst. As ASR-ESAT defines three states per phoneme, we can assign the first state to the closure and the other two states to the burst.
2. Compute the PMFs of this new phone set
3. Now calculate the PLF(k,A) as the mean of the PMF(f) over all phones f whose $A_{ck} = A$.

Of course, constructing the PLF-ESAT is a bit artificial and gives us only an impression of the true phonological features. Nevertheless, with these features we hope to find an answer to the question whether an IPM based on PMFs and PLFs coming from ASR-ESAT can compete with an IPM based on PMFs and PLFs coming from two different recognizers.

Repeating the PLF computation for all phonological features and for two values $A = 1$ and $A = 0$ of each feature results in 48 PLFs per ASR.

3.3. Intelligibility Prediction Model (IPM)

The final step is the conversion of the speaker features into an objective intelligibility score for the speaker. For that purpose we use a regression model that is trained on pathological as well as normal speakers.

3.3.1. Model choice

A variety of statistical learners is available for optimizing regression problems. However, given the fact that the number of features is high compared to the number of speakers, a linear regression model in terms of selected features, possibly in combination with some ad hoc transformation of these features, is about the most complex model we can construct.

3.3.2. Model training

In this study we investigate linear regression models that were based on different subsets selected from the available feature sets WA, LLR, PMF and PLF. A five-fold cross-validation (CV) method is used to identify the feature subset yielding the best performance. The Pearson correlation coefficient (PCC) between the computed and the perceptual intelligibilities is used as the performance measure. To restrict the bias of the CV-approach as much as possible, we have forced the feature selection process to select the same feature subset in all 5 folds of one CV-trial.

Due to the large number of features, it would be impractical to use an exhaustive search for the best subset. Instead, sub-optimal sequential feature selection procedures such as adding or removing 1 feature at the time are used as simple alternatives. Adding or removing more than 2 or 3 features at the time yielded very similar performances.

4. Results and discussion

4.1. General results

In this section we present the performances of several intelligibility prediction models which were built by already confining the feature set from which they can select the features they need.

By only considering the WA features, we obtained PCCs which were low (0.361 for WA-ELIS, only global WA selected) to moderate (0.724 for WA-ESAT, global WA and WA of list A selected). Using the LLR as an alternative to WA we hoped to circumvent the effects of discretization. Unfortunately, the use of this measure leads to PCCs of 0.350 for LLR-ELIS (only global LLR selected) and 0.593 for LLR-ESAT (global LLR and that of list B selected).

The PCCs obtained with models having access to the phonemic or phonological feature sets of one of the ASRs are listed in Table 1. All feature sets yield a very similar performance, be it that the ELIS-based IPMs are more complex.

In the hope to further improve the performance, different combinations of two of these four feature sets were examined. Combining the two feature sets emerging from the same ASR was not expected to yield a very significant improvement since the two feature sets are then derived from the same underlying acoustic models. This is confirmed by the results: combining the ESAT feature sets leads to a PCC of 0.795 (16 selected features), combining the ELIS features sets yields a PCC of 0.815 (45 selected features). More interesting would be to combine the PMF-ESAT and the PLF-ELIS features since they emerge from systems comprising different acoustic models and different alignment strategies. This combination leads to a PCC of 0.858 on the basis of 34 selected features. The scatter plot of the subjective and objective intelligibility scores for this system is shown in Figure 1.

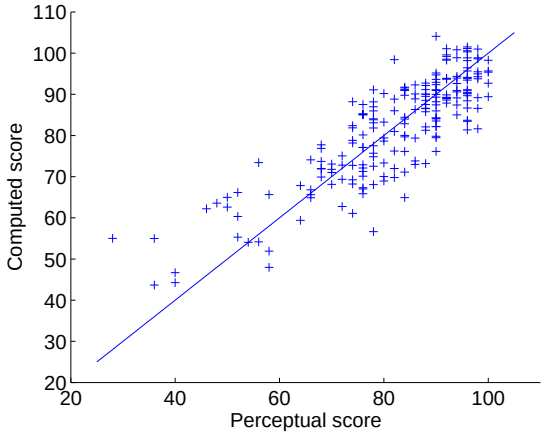

Figure 1: *Computed versus perceptual intelligibility score when combining PMF-ESAT and PLF-ELIS.*

We also created an IPM that could choose its features from the full set of 207 features. This model yielded a PCC of 0.866, which is not significantly better than the 0.858 we got already.

4.2. Pathology-specific intelligibility prediction models

If a clinician is mainly working with one pathology, he is probably more interested in an intelligibility prediction model that is specialized for that pathology. This can be done by selecting the input features yielding the highest PCC between the subjective and objective scores of the test speakers sharing that pathology. We have trained such models for the pathologies dysarthria (DYS), laryngectomy (LARYNX) and hearing impairment (HEAR). The PCCs measured for these models are listed in Table 2. The scatter plot of the computed versus perceptual intelligibility scores emerging from the dysarthria model is shown in Figure 2. The dysarthric speakers are close to the diagonal, but the dispersion of other speakers is clearly increased. The PCCs we obtain compare favorably to the PCCs of 0.88 (for tracheo-oesaphagal speakers) and 0.92 (for speakers with cancer of the oral cavity) reported by Riedhammer et al [4]. Obviously, a direct comparison is difficult to make since in [4] the perceptual intelligibility was just an impression of intelligibility (rated on a 7-point Likert-scale), the populations were smaller than in our study and the intelligibility prediction was evaluated using a leave-one-out paradigm. According to [12], leave-one out produces a higher variance and a higher positive

Table 1: *Pearson correlation coefficients (PCC) for different phonemic and phonological feature sets. N denotes the number of selected features that yielded these results.*

	PMF-ESAT	PMF-ELIS	PLF-ESAT	PLF-ELIS
PCC	0.798	0.737	0.753	0.779
N	15	27	12	23

Table 2: *Pearson correlation coefficients (PCC) for pathology specific IPM (labels are explained in text). N denotes the number of selected features that yielded these result.*

	DYS	LARYNX	HEAR
PCC	0.943	0.907	0.972
N	34	22	46

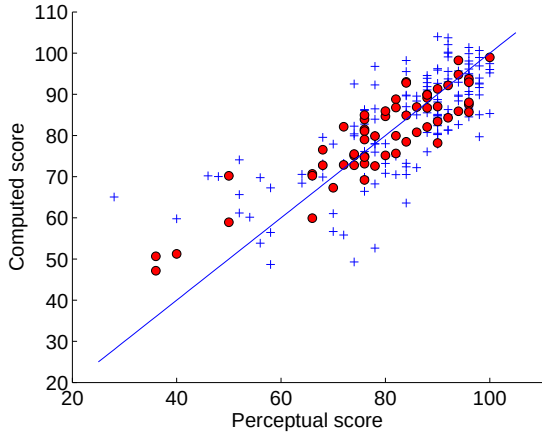

Figure 2: *Computed versus perceptual intelligibility scores emerging from the PMF+PLF intelligibility model of dysarthric speakers. The circles denote the dysarthric speakers, the crosses denote other speakers.*

performance bias than a five-fold CV test.

The largest deviations between computed and subjective intelligibilities are observed for the speakers with a low intelligibility rate. This is a logical consequence of the fact that we only have scarce data in that area. We were not able to record many of such speakers because they often have other disabilities as well and are therefore incapable of performing the test.

4.3. Most relevant features

In order to investigate the clinical plausibility of our approach we have also identified the features with the largest impact on the speech intelligibility. Just selecting the features that contributed to our best IPM would be too restrictive since there are a manifold of alternative feature sets of the same size that would have yielded a very comparable performance. Instead, we selected all feature sequences of length 5 which led to a PCC $>$ 0.68 and we recorded the 15 features appearing most frequently. They are listed in descending order of frequency in Table 3.

To find out if the selected features correspond to important characteristics that have already been described in the literature, an in-depth study of this matter was conducted [13]. It confirmed that the features selected by our model can indeed be linked to characteristics that have been previously associated with pathological speech.

5. Conclusions and future work

In this paper, a first approach toward an automatic assessment of the intelligibility and articulation deficiency of pathological speakers is described. It is shown that alignment-based methods combining phonemic and phonological features yield a correlation between the subjective (human) scores and the objective (computed) scores of about 0.86 for a general model and over 0.90 for a pathology specific model. The correlations for specifc pathologies compete with the interrater agreements measured for perceptual intelligibility assessment. The fact that intelligibility is well predicted by the computed features opens up new possibilities for a more profound articulatory assessment and possibly a better therapy for patients with speech disorders.

Table 3: *Most frequently selected features in decreasing order of frequency (from top left to bottom right).*

/i/	/s/	/z/	/2/	/O/
/l/	/j/	Not lateral	Lateral	/A/
/x/	/A+/	Mid	/o/	Low

6. Acknowledgements

This work was supported by the Flemish Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) (contract SBO/40102).

7. References

[1] K. M. Yorkston, E. A. Strand, and M. Kennedy, "Comprehensibility of dysarthric speech: implications for assessment and treatment planning," *American Journal of Speech-Language-Pathology*, vol. 5, pp. 55–66, 1996.

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

[3] 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.

[4] K. Riedhammer, G. Stemmer, T. Haderlein, M. Schuster, F. Rosanowski, E. Nöth., and A. Maier, "Towards robust automatic evaluation of pathologic telephone speech," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2007, pp. 717–722.

[5] A. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke, "Automatic scoring of the intelligibility in patients with cancer of the oral cavity," in *Interspeech*, 2007, pp. 1206–1209.

[6] 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*, In Press.

[7] 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.

[8] K. Demuynck, "Extracting, modelling and combining information in speech recognition," Ph.D. dissertation, K.U.Leuven, ESAT, 2001.

[9] F. Stouten and J. Martens, "On the use of phonological features for pronunciation scoring," in *International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. 329–332.

[10] ——, "Speech recognition with phonological features: some issues to attend," in *Interspeech*, 2006, pp. 357–360.

[11] J. Carmichael and P. Green, "Revisiting dysarthria assessment intelligibility metrics," in *8th International Conference on Spoken Language Processing (ICSLP) October 4-8, Korea*, 2004, pp. 742–745.

[12] J. F. T. Hastie, R. Tibshirani, *The Elements of Statistical Learning*. Springer, 2001.

[13] 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*, In Press.