Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 629030, 9 pages
doi:10.1155/2009/629030

Research Article

Automated Intelligibility Assessment of Pathological Speech Using Phonological Features

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

[1] *Department of Electronics and Information Systems, Ghent University, 9000 Ghent, Belgium*
[2] *Antwerp University Hospital, University of Antwerp, 2650 Edegem, Belgium*

Correspondence should be addressed to Catherine Middag, catherine.middag@ugent.be

Received 31 October 2008; Accepted 24 March 2009

Recommended by Juan I. Godino-Llorente

It is commonly acknowledged that word or phoneme intelligibility is an important criterion in the assessment of the communication efficiency of a pathological speaker. People have therefore put a lot of effort in the design of perceptual intelligibility rating tests. These tests usually have the drawback that they employ unnatural speech material (e.g., nonsense words) and that they cannot fully exclude errors due to listener bias. Therefore, there is a growing interest in the application of objective automatic speech recognition technology to automate the intelligibility assessment. Current research is headed towards the design of automated methods which can be shown to produce ratings that correspond well with those emerging from a well-designed and well-performed perceptual test. In this paper, a novel methodology that is built on previous work (Middag et al., 2008) is presented. It utilizes phonological features, automatic speech alignment based on acoustic models that were trained on normal speech, context-dependent speaker feature extraction, and intelligibility prediction based on a small model that can be trained on pathological speech samples. The experimental evaluation of the new system reveals that the root mean squared error of the discrepancies between perceived and computed intelligibilities can be as low as 8 on a scale of 0 to 100.

1. Introduction

In clinical practice there is a great demand for fast and reliable methods for assessing the communication efficiency of a person with a (pathological) speech disorder. It is argued in several studies (e.g., [1]) that intelligibility is an important criterion in this assessment. Therefore several perceptual tests aiming at the measurement of speech intelligibility have been conceived [2–4]. One of the primary prerequisites for getting reliable scores is that the test should be designed in such a way that the listener cannot guess the correct answer based solely on contextual information. That is why these tests use random word lists, varying lists at different trials, real words as well as pseudowords, and so forth. Another important issue is that the listener should not be too familiar with the tested speaker since this creates a positive bias. Finally, if one wants to use the test for monitoring the efficiency of a therapy, one cannot work with the same listener all the time because this would introduce a bias shift. The latter actually excludes the speaker's therapist as a listener, which is very unfortunate from a practical viewpoint.

For the last couple of years there has been a growing interest in trying to apply automatic speech recognition (ASR) for the automation of the traditional perceptual tests [5–8]. By definition an ASR is an unbiased listener, but is it already reliable enough to give rise to computed intelligibility scores that correlate well with the scores obtained from a well-designed and well-performed perceptual test? In this paper, we present and evaluate an automated test which seems to provide such scores.

The simplest approach to automated testing is to let an ASR listen to the speech, to let it perform a lexical decoding of that speech, and to compute the intelligibility as the percentage of correctly decoded words or phonemes. Recent work [9, 10] has demonstrated that in case the sketched approach is applied to read text passages of speakers with a particular disorder (e.g., dysarthric speakers or laryngectomies) it can yield intelligibilities that correlate well with an impression of intelligibility, expressed on a 7-point Likert scale [11].

In order to explore the potential of the approach in more demanding situations, we have let a state-of-the-art ASR system [12] recognize isolated monosyllabic words and pseudowords spoken by a variety of pathological speakers (different types of pathology and different degrees of severity of that pathology). The perceptual intelligibilities against which we compared the computed ones represented intelligibility at the phone level. The outcome of our experiments was that the correlations between the perceptual and the computed scores were only moderate [13]. This is inline with our expectations since the ASR employs acoustic models that were trained on the speech of nonpathological speakers. Consequently, when confronted with severely disordered speech, the ASR is asked to score the sounds that are in many respects very different from the sounds it was trained on. This means that acoustic models are asked to make extrapolations in areas of the acoustic space that were not examined at all during training. One cannot expect that under these circumstances a lower acoustic likelihood always points to a larger deviation (distortion) of the observed pronunciation from the norm.

Based on this last argument we have conceived an alternative approach. It first of all employs phonological features as an intermediate description of the speech sounds. Furthermore, it computes a series of features used for characterizing the voice of a speaker, and it employs a separate intelligibility prediction model (IPM) to convert these features into a computed intelligibility. Our first hypothesis was that even in the case of severe speech disorders, some of the articulatory dimensions of a sound may still be more or less preserved. A description of the sounds in an articulatory feature space may possibly offer a foundation for at least assessing the severity of the relatively limited distortions in these articulatory dimensions. Note that the term “articulatory” is usually reserved to designate features stemming from direct measurements of articulatory movements (e.g., by means of an articulograph). We adopt the term “phonological” for features that are also intended to describe articulatory phenomena, although here they are derived from the waveform. Our second hypothesis was that it would take only a simple IPM with a small number of free parameters to convert the speaker features into an intelligibility score, and therefore that this IPM can be trained on a small collection of both pathological and normal speakers.

We formerly developed an initial version of our system [13], and we were able to demonstrate that its computed intelligibilities correlated well with perceived phone-level intelligibilities [14] for our speech material. However, these good correlations could only be attained with a system incorporating two distinct ASR components: one working directly in the acoustic feature space and one working in the phonological feature space. In this paper we present significant improvements of the phonological component of our system, and we show that as a result of these improvements we can now obtain high accuracy using phonological features alone. This means that we now obtain good results with a much simpler system comprising only one ASR comprising no more than 55 context-independent acoustic models.

The rest of this paper is organized as follows. In Section 2, we briefly describe the perceptual test that was automated and the pathological speech corpus that was available for the training and evaluation of our system. In Section 3 we present the system architecture, and we briefly discuss the basic operations performed by the initial stages of the system. The novel speaker feature extractor and the training of the IPM are discussed in Sections 4 and 5, respectively. In Section 6 we assess the reliability of the new system and compare it to that of the original system. The paper ends with a conclusion and some directions for future work.

2. Perceptual Test and Evaluation Database

The subjective test we have automated is the Dutch Intelligibility Assessment (DIA) test [4], one which was specifically designed with the aim to measure the intelligibility of Dutch speech at the phoneme level. Each speaker reads 50 consonant-vowel-consonant (CVC) words but with one relaxation, namely, those words with one of the two consonants missing are also allowed. The words are selected from three lists: list A is intended for testing the consonants in a word initial position (19 words including one with a missing initial consonant), list B is intended for testing them in a word final position (15 words including one with a missing final consonant), and list C is intended for testing the vowels and diphthongs in a word central position (16 words with an initial and final consonant). To avoid guessing by the listener, there are 25 variants of each list, and each variant contains existing words as well as pronounceable pseudowords. For each test word, the listener must complete a word frame by filling in the missing phoneme or by indicating the absence of that phoneme. In case the initial consonant is tested, the word frame could be something like “.it” or “.ol”. The perceptual intelligibility score is calculated as the percentage of correctly identified phonemes. Previous research [4, 15] has demonstrated that the intelligibility scores derived from the DIA are highly reliable (an interrater correlation of 0.91 and an intrarater correlation of 0.93 [15]).

In order to train and test our automatic intelligibility measurement system, we could dispose of a corpus of recordings from 211 speakers. All speakers uttered 50 CVC words (the DIA test) and a short text passage.

The speakers belong to 7 distinct categories: 51 speakers without any known speech impairment (the control group), 60 dysarthric speakers, 12 children with cleft lip or palate, 42 persons with pathological speech secondary to hearing impairment, 37 laryngectomized speakers, 7 persons diagnosed with dysphonia, and 2 persons with a glossectomy.

The DIA recordings of all speakers were scored by one trained speech therapist. This therapist was however not familiar with the recorded patients. The perceptual (subjective) phoneme intelligibilities of the pathological training speakers range from 28 to 100 percent with a mean of 78.7 percent. The perceptual scores of the control speakers range from 84 to 100 percent, with a mean of 93.3 percent. More details on the recording conditions and the severity of the speech disorders can be found in [14].

We intend to make the data freely available for research through the Dutch Speech and Language Resources agency (TST-centrale), but this requires good documentation in English first. In the meantime, the data can already be obtained by simple request (just contact the first author of this paper).

3. An Automatic Intelligibility Measurement System

As already mentioned in the introduction, we have conceived a new speech intelligibility measurement system that is more than just a standard word recognizer. The architecture of the system is depicted in Figure 1. The acoustic front-end extracts a stream of mel-frequency cepstral coefficients (MFCC) [16] feature vectors from the waveform. At every time $t = 1, \ldots, T$ which is a multiple of 10 milliseconds, it computes a vector X_t of 12 MFCCs plus a log-energy (all derived from a segment of 30 milliseconds centered around t). This MFCC feature stream is then converted into a phonological feature stream. At each time t, the phonological feature detector computes a vector Y_t of 24 components each representing the posterior probability $P(A_i \mid X_{t-5}, \ldots, X_{t+5})$ that one of 24 binary phonological classes A_i ($i = 1, \ldots, 24$) is "supported by the acoustics" in a 110 milliseconds window around time t. The full list of phonological classes can be found in [17]. Some typical examples are the classes *voiced* (= vocal source class), *burst* (= manner class), *labial* (= place-consonant class), and *mid-low* (= vowel class). The phonological feature detector is a conglomerate of four artificial neural networks that were trained on continuous speech uttered by normal speakers [17].

The forced alignment system lines up the phonological feature stream with a typical (canonical) acoustic-phonetic transcription of the target word. This transcription is a sequence of basic acoustic-phonetic units, commonly referred to as phones [18]. The acoustic-phonetic transcription is modeled by a sequential finite state machine composed of one state per phone. The states are context-independent, meaning that all occurrences of a particular phone are modeled by the same state. This is considered acceptable because coarticulations can be handled in an implicit way by the phonological feature detector. In fact, the latter analyzes a long time interval for any given timeframe, and this window can expose most of the contextual effects. Each state is characterized by a set of canonical values A_{ci} for the phonological classes A_i. These values can either be 1 (= on, present), 0 (= off, absent), or *irrelevant* (= both values are equally acceptable). Self-loops and skip transitions make it possible to handle variable phone durations and phone omissions in an easy way.

The alignment system is instructed to return the state sequence $S = \{s_1, \ldots, s_T\}$ with the largest posterior probability $P(S \mid X_1, \ldots, X_T)$. This probability is approximated as follows (see [17] for more details):

$$P(S \mid X_1, \ldots, X_T) = \prod_{t=1}^{T} P(s_t \mid X_{t-5}, \ldots, X_{t+5}) \frac{P(s_t \mid s_{t-1})}{P(s_t)},$$
$$P(s_t \mid X_{t-5}, \ldots, X_{t+5}) = \left[\prod_{A_{ci}(s_t)=1} Y_{ti} \right]^{1/N_p(s_t)}, \tag{1}$$

with $N_p(s_t)$ representing the number of classes with a positive canonical value for state s_t. The transition probabilities $P(s_t \mid s_{t-1})$ and the prior state probabilities $P(s_t)$ were trained on normal speech. The probability $P(s_t \mid X_{t-5}, \ldots, X_{t+5})$ is hereafter shortnoted as $P(s_t \mid X_t)$.

Once the 3 tuples $(s_t, Y_t, P(s_t \mid X_t))$ are available for all frames of all utterances of one speaker, the speaker feature extractor can derive from these 3 tuples (and from the canonical values of the phonological classes in the different states) a set of phonological features that characterize the speaker. The Intelligibility Prediction Model (IPM) then converts these speaker features into a computed phoneme intelligibility score.

In the subsequent sections, we will provide a more detailed description of the last two processing stages since these are the stages that mostly distinguish the new from the original system.

4. Speaker Feature Extraction

In [13], only context-independent speaker features were derived from the alignments. In this work we will benefit from the binary nature of the phonological classes to identify an additional set of context-dependent speaker features that can be extracted from these alignments.

The extraction of speaker features is always based on averaging either $P(s_t \mid X_t)$ or Y_t over frames that were assigned to a particular state or set of states. The averaging is not restricted to frames that, according to the alignment, contribute to the realization of a phoneme that is being tested in the DIA (e.g., the initial consonant of the word). We let the full utterances and the corresponding state sequences contribute to the feature computation because we assume that this should lead to a more reliable (stable) characterization of the speaker. However, at certain places, we have compensated for the fact that not every speaker has pronounced the same words (due to subtest variants), and therefore, that the distribution of phonemes can differ from speaker to speaker as well.

4.1. Phonemic Features (PMFs). A phonemic feature PMF(f) for phone f is derived as the mean of $P(s_t|X_t)$ over all frames X_t that were assigned to a state s_t which is equal to f (there is 1 state per phone). Repeating this for every phone in the inventory then gives rise to 55 PMFs of the form

$$\text{PMF}(f) = \langle P(s_t \mid X_t) \rangle_{t; s_t = f} \quad f = 1, \ldots, 55, \tag{2}$$

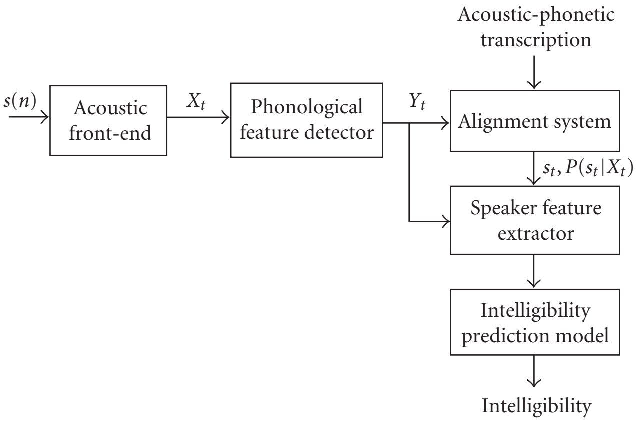

Acoustic-phonetic transcription
$s(n)$
Acoustic front-end
X_t
Phonological feature detector
Y_t
Alignment system
$s_t, P(s_t|X_t)$
Speaker feature extractor
Intelligibility prediction model
Intelligibility

Figure 1: Architecture of the automatic intelligibility measurement system.

with $\langle x\rangle_{\text{selection}}$ representing the mean of x over the frames specified by the selection.

4.2. Phonological Features (PLFs). Instead of averaging the posterior probabilities $P(s_t \mid X_t)$, one can also average the phonological features Y_{ti} ($i = 1,\dots,24$). In particular, one can take the mean of Y_{ti} (for some i) over all frames that were assigned to one of the phones that are characterized by a canonical value $A_{ci} = A$ for feature class A_i (A can be either 1 or 0 here). This mean score is thus generally determined by the realizations of multiple phones. Consequently, since different speakers have uttered different word lists, the different phones could have a speaker-dependent weight in the computed means. In order to avoid this, the simple averaging scheme is replaced by the following two-stage procedure:

(1) take the mean of Y_{ti} over all frames that were assigned to a phone f whose $A_{ci}(f) = A$, denote this mean as PLF(f,i,A), and repeat the procedure for all valid combinations (f,i,A);

(2) compute PLF(i,A) as the mean over f of the PLF(f,i,A) that were obtained in the previous stage.

This procedure gives equal weights to every phone contributing to PLF(i,A). Written in mathematical notation, one gets

$$\text{PLF}(f,i,A) = \langle Y_{ti}\rangle_{t;s_t=f;A_{ci}(f)=A} \quad \forall \text{ valid } (f,i,A),$$

$$\text{PLF}(i,A) = \langle \text{PLF}(f,i,A)\rangle_{f;A_{ci}(f)=A} \quad i = 1,\dots,24;\ A = 0,1. \tag{3}$$

Since for every of the 24 phonological feature classes there are phones with canonical values 0 and 1 for that class, one always obtains 48 phonological features. The 24 phonological features PLF$(i,1)$ are called positive features because they measure to what extent a phonological class that was supposed to be present during the realization of certain phones is actually supported by the acoustics observed during these realizations. The 24 phonological features PLF$(i,0)$ are called negative features. We add this negative PLF set because it is important for a patient's intelligibility not only that phonological features occur at the right time but also that they are absent when they should be.

4.3. Context-Dependent Phonological Features (CD-PLFs). It can be expected that pathological speakers encounter more problems with the realization of a particular phonological class in some contexts than in others. Consequently it makes sense to compute the mean value of a phonological feature Y_{ti} under different circumstances that take not only the canonical value of feature class A_i in the tested phone into account but also the properties of the surrounding phones. Since the phonological classes are supposed to refer to different dimensions of articulation, it makes sense to consider them more or less independently, and therefore, to consider only the canonical values of the tested phonological class in these phones as context information. Due to the ternary nature of the phonological class values (on, off, irelevant), the number of potential contexts per (i,A) is limited to $3\times3 = 9$. If we further include "silence" as a special context to indicate that there is no preceding or succeeding phone, the final number of contexts is 16. Taking into account that PLFs are only generated for canonical values A of 0 and 1 (and not for irrelevant), the total number of sequences of canonical values (SCVs) for which to compute a CD-PLF is $24 \times 2 \times 16 = 768$. This number is however an upper bound since many of these SCVs will not occur in the 50 word utterances of the speaker.

In order to determine in advance all the SCVs that are worthwhile to consider in our system, we examined the canonical acoustic-phonetic transcriptions of the words in the different variants of the A, B, or C-lists, respectively. We derived from these lists how many times they contain a particular SCV. We then retained only those SCVs that appeared at least twice in any combination of variants one could make. It is easy to determine the minimal number of occurences of each SCV. One just needs to determine the number of times each variant of the A-list contains the SCV and to record the minimum over these times to get an A-count. Similarly one determines a B and a C-counts, and one takes the sum of these counts. For our test, we found that 123 of the 768 SCVs met the condition we set out.

If A^L and A^R represent the canonical values of feature class A_i in the left and right context phone, the computation of a context-dependent feature for the combination (A, A^L, A^R) is obtained by means of a two-stage scheme:

(1) take the mean of Y_{ti} over all frames which were assigned to a phone f having a canonical value $A_{ci}(f) = A$ (A can be either 1 or 0 here) and appearing between phones whose canonical values of class A_i are A^L and A^R, denote this mean as $\text{PLF}(f, i, A, A^L, A^R)$, and repeat the procedure for all combinations (f, i, A, A^L, A^R) occurring in the data,

(2) compute $\text{PLF}(i, A, A^L, A^R)$ as the mean over f of the $\text{PLF}(f, i, A, A^L, A^R)$ that were computed in the first stage.

Again, this procedure gives equal weights to all the phones that contribute to a certain CD-PLF. In mathematical notation one obtains

$$\begin{aligned}
&\text{PLF}\left(f, i, A, A^L, A^R\right) \\
&\quad = \langle Y_{ti} \rangle_{t;\, s_t = f;\, A_{ci}=A;\, A^L_{ci}=A^L;\, A^R_{ci}=A^R} \\
&\qquad \forall \text{ occurring } \left(f, i, A, A_L, A^R\right), \\
&\text{PLF}\left(i, A, A^L, A^R\right) \\
&\quad = \left\langle PLF\ (f, i, A, A^L, A^R)\right\rangle_{f;\, \text{occurring } (f,i,A,A^L,A^R)} \\
&\qquad\qquad \forall \text{ occurring } \left(i, A, A^L, A^R\right),
\end{aligned} \tag{4}$$

with A_{ci}, A^L_{ci}, and A^R_{ci} being short notations for, respectively, the canonical values of A_i in the state visited at time t, in the state from where this state was reached at some time before t, and in the state which is visited after having left the present state at some time after t.

Note that the context is derived from the phone sequence that was actually realized according to the alignment system. Consequently, if a phone is omitted, a context that was not expected from the canonical transcriptions can occur, and vice versa. Furthermore, there may be fewer observations than expected for the SCV that has the omitted phone in central position. In the case that no observation of a particular SCV would be available, the corresponding feature is replaced by its expected value (as derived from a set of recorded tests).

5. Intelligibility Prediction Model (IPM)

When all speaker features are computed, they need to be converted into an objective intelligibility score for the speaker. In doing so we use a regression model that is trained on both pathological and normal speakers.

5.1. Model Choice. A variety of statistical learners is available for optimizing regression problems. However, in order to avoid overfitting, only a few of these can be applied to our data set. This is because the number of training speakers (211) is limited compared to the number of features (e.g., 123 CD-PLFs) per speaker. A linear regression model in terms of selected features, with the possible combination of some ad hoc transformation of these features, is about the most complex model we can construct.

5.2. Model Training. We build linear regression models for different feature sets, namely, PMF, PLF, and CD-PLF, and combinations thereof. A fivefold cross-validation (CV) method is used to identify the feature subset yielding the best performance. In contrast to our previous work, we no longer take the Pearson Correlation Coefficient (PCC) as the primary performance criterion. Instead, we opt for the root mean squared error (RMSE) of the discrepancies between the computed and the measured intelligibilities. Our main arguments for this change of strategy are the following.

First of all, the RMSE is directly interpretable. In case the discrepancies (errors) are normally distributed, 67% of the computed scores lie closer than the RMSE to the measured (correct) scores. Using the Lilliefors test [19] we verified that, in practically all the experiments we performed, the errors were indeed normally distributed.

A second argument is that we want the computed scores to approximate the correct scores directly. Per test set, the PCC actually quantifies the degree of correlation between the correct scores and the best linear transformation of the computed scores. As this transformation is optimized for the considered test set, the PCC may yield an overly optimistic evaluation result.

Finally, we noticed that if a model is designed to cover a large intelligibility range, and if it is evaluated on a subgroup (e.g., the control group) covering only a small subrange, the PCC can be quite low for this subgroup even though the errors remain acceptable. This happens when the rankings of the speakers of this group along the perceptual and the objective scores, respectively, are significantly different. The RMSE results were found to be much more stable across subgroups.

Due to the large number of features, an exhaustive search for the best subset would take a lot of computation time. Therefore we investigated two much faster but definitely suboptimal sequential procedures. The so-called forward procedure starts with the best combination of 3 features and adds one feature (the best) at the time. The so-called backward procedure starts with all the features and removes one feature at the time.

Figure 2 illustrates a typical variation of RMSE versus the number of features being selected. By measuring not only the global RMSE but also the individual RMSEs in the 5 folds of the CV-test, one can get an estimate of the standard deviation on the global RMSE for a particular selected feature set. In order to avoid that too many features are being selected we have adopted the following 2-step procedure: (1) determine the selected feature set yielding the minimal RMSE; (2) select the smallest feature set yielding an RMSE that is not larger than the minimal (best) RMSE augmented with the estimated standard deviation on that RMSE.

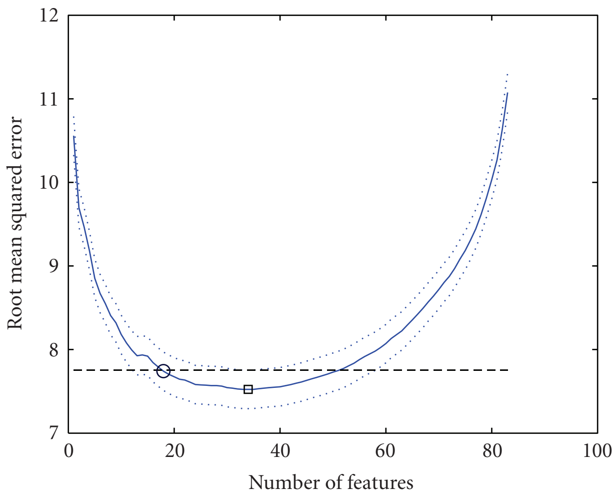


Figure 2: Typical evolution of the root mean squared error (RMSE) as a function of the number of selected features for the forward selection procedure. Also indicated is the evolution of RMSE ± σ with σ representing the standard deviation on the RMSEs found for the 5 folds. The square and the circle show the sizes of the best and the actually selected feature subsets.

6. Results and Discussion

We present results for the new system as well as for a previously published system that was much more complex since it comprised two subsystems each containing a different ASR and each generating a set of speaker features. The first subsystem generated 55 phonemic features (PMF-tri) originating from acoustic scores computed by state-of-the-art triphone acoustic models in the MFCC feature space. The second subsystem generated 48 phonological features (PLFs) in the way described in Section 4.2. The speaker features of the two subsystems could be combined before they were supplied to the intelligibility prediction model.

6.1. General Results. We have used the RMSE criterion to obtain three general IPMs (trained on all speakers) that were based on the speaker features generated by our original system. The first model only used the phonemic features (PMF-tri) emerging from the first subsystem, the second one applied the phonological features (PLF) emerging from the other subsystem, and the third one utilized the union of these two feature sets (PMF-tri + PLF). The number of selected features and the RMSEs for these models are listed in the first three rows of Table 1.

Next, we examined all the combinations of 1, 2, or 3 speaker feature sets as they emerged from the new system. The figures in Table 1 show that all IPMs using the CD-PLFs perform the same as our previous best system: PMF-tri + PLF. In the future as we look further into underlying articulatory problems of pathological speakers, it will be most pertinent to opt for an IPM based solely on articulatory information such as PLF + CD-PLF.

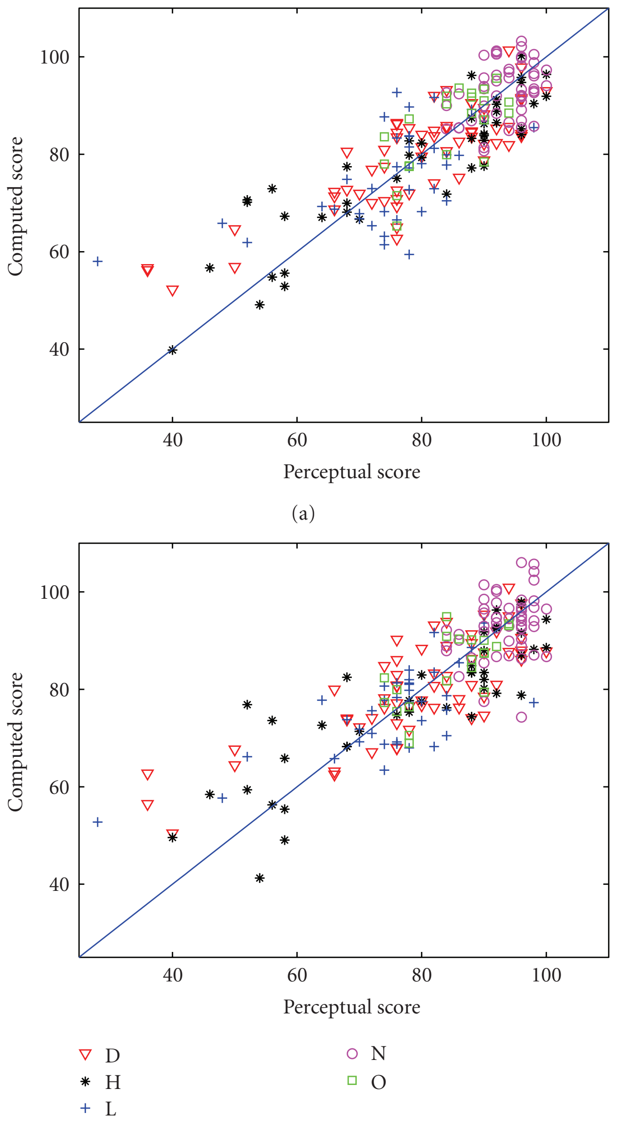
Taking this IPM as our reference system, the Wilcoxon singed-rank test [19] has revealed the following: (1) there is no significant difference between the accuracy of the new reference system and that of the formerly published system, (2) the context-dependent feature set yields a significantly better accuracy than any of the context-independent feature sets, (3) the addition of context-independent features to CD-PLF only yields a nonsignificant improvement, and (4) a combination of context-independent phonemic and phonological features emerging from one ASR (PMF + PLF) cannot compete with a combination of similar features (PMF-tri + PLF) originating from two different ASRs. Although maybe a bit disappointing at first glance, the first conclusion is an important one because it shows that the new system with only one ASR comprising 55 context-independent acoustic states achieves the same performance as our formerly published system with two ASRs, one of which is a rather complex one comprising about thousand triphone acoustic states.

Table 1: Number of selected features and RMSE for a number of general models (trained on all speakers) created for different speaker feature sets. The features with suffix “tri” emerge from our previously published system. Results differing significantly from the ones of our reference system PLF + CD-PLF are marked in bold.

Speaker features	Selected features	RMSE
PMF-tri	5	**8.9**
PLF	16	**9.2**
PMF-tri + PLF	19	7.7
PMF	11	**10.1**
PLF	16	**9.2**
CD-PLF	21	8.2
PLF + CD-PLF	27	7.9
PMF + CD-PLF	31	7.8
PMF + PLF	20	**9.0**
PMF + PLF + CD-PLF	42	7.8

Table 2: Root mean squared error (RMSE) for pathology specific IPMs (labels are explained in the text) based on several speaker feature sets. N denotes the number of selected features. The results which differ significantly from the reference system PLF + CD-PLF are marked in bold.

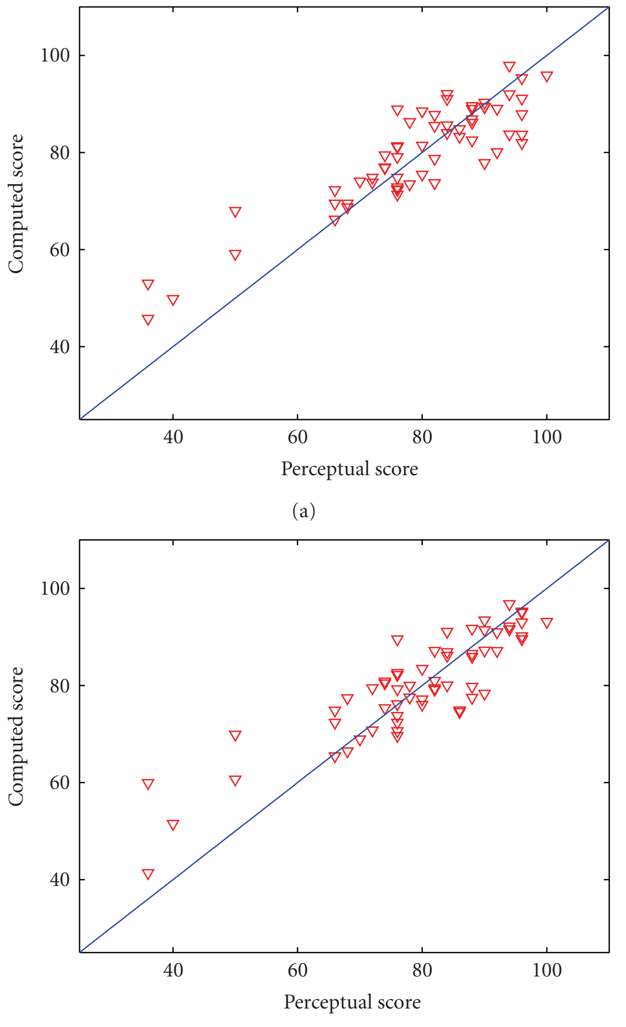
		DYS	LARYNX	HEAR
CD-PLF	RMSE	6.4	5.2	5.8
	N	28	19	43
PMF + PLF	RMSE	**7.9**	**7.3**	**8.1**
	N	12	8	22
PMF + CD-PLF	RMSE	6.1	4.3	**3.9**
	N	22	31	55
PLF + CD-PLF	RMSE	*6.1*	*5.3*	*4.8*
	N	28	17	52
PMF + PLF + CD-PLF	RMSE	5.9	4.1	**4.2**
	N	38	28	49
PMF-tri + PLF	RMSE	6.4	7.6	5.5
	N	26	10	22



(a)

(b)

Figure 3: Computed versus perceptual intelligibility scores emerging from the systems PMF-tri + PLF (a) and PLF + CD-PLF (b). Different symbols were used for dysarthric speakers (D), persons with hearing impairment (H), laryngectomized speakers (L), speakers with normal speech (N), and others (O).



(a)

(b)

Figure 4: Computed versus perceptual intelligibility scores emerging from the PMF-tri + PLF (a) and PLF + CD-PLF (b) for dysarthric speakers.

Scatter plots of the subjective versus the objective intelligibility scores for the systems PMF-tri + PLF and PLF + CD-PLF are shown in Figure 3. They confirm that most of the dots are in vertical direction less than the RMSE (about 8 points) away from the diagonal which represents the ideal model. They also confirm that the RMSE emerging from our former system is slightly lower than that emerging from our new system.

The largest deviations from the diagonal appear for the speakers with a low intelligibility rate. This is a logical consequence of the fact that we only have a few such speakers in the database. This means that the trained IPM will be more specialized in rating medium to high-quality speakers. Consequently, it will tend to produce overrated intelligibilities for bad speakers. We were not able to record many more bad speakers because they often have other disabilities as well and are therefore incapable of performing the test. By giving more weight to the speakers with low perceptual scores during the training of the IPM, it is possible to reduce the errors for the low perceptual scores at the expense of only a small increase of the RMSE caused by the slightly larger errors for the high perceptual scores.

6.2. 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 in that pathology. Our hypothesis is that since people with different pathologies are bound to have different articulation problems, pathology specific models should select pathology-specific features. We therefore search for the feature set offering the lowest RMSE on the speakers of the validation group with the targeted pathology. However, for training the regression coefficients of the IPM we use all the speakers in the training fold. This way we can alleviate the problem of having an insufficient number of pathology-specific speakers to compute reliable regression coefficients. The characteristics of the specialized models for dysarthria (DYS), laryngectomy (LARYNX), and hearing impairment (HEAR) can be found in Table 2. The results which differ significantly from the reference results are marked in bold; the reference results are themselves marked in italic. The data basically support the conclusions that were drawn from Table 1, with two exceptions: (1) for the HEAR model, adding PMF to CD-PLF turns out to yield a significant improvement now, and (2) for the LARYNX model, the combination PMF + PLF is not significantly worse than PMF-tri + PLF.

Scatter plots of the computed versus the perceptual intelligibility scores emerging from the former (PMF-tri + PLF) and the new (PLF + CD-PLF) dysarthria model are shown in Figure 4.

In [13] we already compared results obtained with our former system to results reported by Riedhammer et al. [9] for a system also comprising two state-of-the-art ASR systems. Although a direct comparison is difficult to make, it appears that our results emerging from an evaluation on a diverse speaker set are very comparable to those reported in [9], even though the latter emerged from an evaluation on a narrower set of speakers (either tracheo-oesaphagal speakers or speakers with cancer of the oral cavity).

7. Conclusions and Future Work

In our previous work [13], we showed that an alignment-based method combining two ASR systems can yield good correlations between subjective (human) and objective (computed) intelligibility scores. For a general model, we obtained Pearson correlations of about 0.86. For a dysarthria specific model these correlations were as large as 0.94. In the present paper we have shown that by introducing context-dependent phonological features it is possible to achieve equal to higher accuracies by means of a system comprising only one ASR which works on phonological features that were extracted from the waveform by a set of neural networks.

Now that we have an intelligibility score which is described in terms of features that refer to articulatory dimensions, we can start to think of extracting more detailed information that can reveal the underlying articulatory problems of a tested speaker.

In terms of technology, we still need to conceive more robust speaker feature selection procedures. We must also examine whether an alignment model remains a viable model for the analysis of severely disordered speech. Finally, we believe that there exist more efficient ways of using the new context-dependent phonological features than the one adopted in this paper (e.g., clustering of contexts, better dealing with effects of phone omissions). Finding such ways should result in further improvements of the intelligibility predictions.

Acknowledgment

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

References

[1] R. D. Kent, Ed., *Intelligibility in Speech Disorders: Theory, Measurement, and Management*, John Benjamins, Philadelphia, Pa, USA, 1992.

[2] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–499, 1989.

[3] R. D. Kent, "The perceptual sensorimotor examination for motor speech disorders," in *Clinical Management of Sensorimotor Speech Disorders*, pp. 27–47, Thieme Medical, New York, NY, USA, 1997.

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

[5] J. Carmichael and P. Green, "Revisiting dysarthria assessment intelligibility metrics," in *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP '04)*, pp. 742–745, Jeju, South Korea, October 2004.

[6] J.-P. Hosom, L. Shriberg, and J. R. Green, "Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (ASR) methods," *Journal of Medical Speech-Language Pathology*, vol. 12, no. 4, pp. 167–171, 2004.

[7] H.-Y. Su, C.-H. Wu, and P.-J. Tsai, "Automatic assessment of articulation disorders using confident unit-based model adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '08)*, pp. 4513–4516, Las Vegas, Nev, USA, March 2008.

[8] P. Vijayalakshmi, M. R. Reddy, and D. O'Shaughnessy, "Assessment of articulatory sub-systems of dysarthric speech using an isolated-style phoneme recognition system," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '06)*, vol. 2, pp. 981–984, Pittsburgh, Pa, USA, September 2006.

[9] K. Riedhammer, G. Stemmer, T. Haderlein, et al., "Towards robust automatic evaluation of pathologic telephone speech," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU '07)*, pp. 717–722, Kyoto, Japan, December 2007.

[10] 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 *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, vol. 1, pp. 1206–1209, Antwerpen, Belgien, August 2007.

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

[12] K. Demuynck, J. Roelens, D. V. Compernolle, and P. Wambacq, "Spraak: an open source speech recognition and automatic annotation kit," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '08)*, pp. 495–496, Brisbane, Australia, September 2008.

[13] 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 (Interspeech '08)*, pp. 1745–1748, Brisbane, Australia, September 2008.

[14] 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.

[15] G. Van Nuffelen, M. De Bodt, C. Guns, F. Wuyts, and P. Van de Heyning, "Reliability and clinical relevance of segmental analysis based on intelligibility assessment," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 5, pp. 264–268, 2008.

[16] 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. 28, no. 4, pp. 357–366, 1980.

[17] F. Stouten and J.-P. Martens, "On the use of phonological features for pronunciation scoring," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 1, pp. 329–332, Toulouse, France, May 2006.

[18] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Darpa timit acoustic-phonetic continuous speech corpus CD-ROM," Tech. Rep. NISTIR 4930, National Institute of Standards and Technology, Gaithersburgh, Md, USA, 1993.

[19] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, CRC Press, Boca Raton, Fla, USA, 2004.