Speech technology based assessment of dysarthric speech: preliminary results

G. Van Nuffelen°*, C. Middag^, J.P. Martens^& M. De Bodt°~
°Antwerp University Hospital Belgium, Department of Oto-rhino-laryngology, Belgium
*University of Antwerp, Department of Medical Sciences, Belgium
^University of Ghent, Department of Electronics and Information Systems (ELIS), Belgium
~University of Ghent, Department of Speech Language Pathology and Audiology, Belgium

Purpose: One of the objectives of the SPACE-project (Speech Algorithms for Clinical and Educational Applications) is to develop a speech technology based clinical assessment that provides reliable quantitative analyses of pathological speech.
Method: Four automatic speech processing systems were applied on monosyllabic word recordings of a standardized Dutch phoneme intelligibility assessment. Systems 1 and 2 are automatic word recognizers which provide word accuracy rates. System 1 (WAR-ACF) was supplied with standard acoustic features, system 2 (WAR-ARF) with articulatory features, derived from the acoustic features. Systems 3 (CS-ACF) and 4 (CS-ARF) are automatic speech aligners that determine the best alignment between a speech sample and its canonical phonetic transcription. Confidence scores (CS) are computed for each phoneme (system 3) or for each articulatory feature (system 4). These CS are finally converted into a global score, designed to maximally agree with the perceptual intelligibility score. Samples of 60 dysarthric speakers were analyzed objectively by the four systems and perceptually by an experienced speech-language-pathologist. Pearson correlation coefficients (r) between the objective and the perceptual intelligibility scores are estimated by means of 5-fold cross validation experiments.
Results: The correlations for systems 1 and 2 were respectively found to be moderate (r:0.56) and low (r:0.33). However, the alignment-based systems resulted in much higher correlations (system 3: r:0.72; system 4: r:0.72).
Conclusions: Alignment-based systems, provide more reliable intelligibility scores than recognition-based systems. No significant difference was found between working with acoustic and working with articulatory features. The current results are encouraging but further refinements are needed.

INTRODUCTION

Dysarthria is a neurological motor speech disorder that is characterized by slow, weak, imprecise, and/or uncoordinated movements of the speech musculature (1). Dysarthria may affect all dimensions of speech, namely articulation, resonance, voice and prosody, resulting in decreased intelligibility. Intelligibility is defined as the accuracy with which a listener is able to decode the acoustic signal of a speaker (2). Since intelligibility can be considered as the product of the four main dimensions of speech, measuring a person's intelligibility is highly relevant in clinical practice. Until now, intelligibility assessments are mainly based on auditory perceptual judgments involving a speaker, a message, a transmission system and a listener (3). Consequently, estimating or measuring intelligibility is a subjective procedure, which has a lot of intrinsic variables. To obtain reliable intelligibility scores listener's variables like familiarity with the test items, predictability of the test items and familiarity with the speaker and/or speech pathology must be controlled. Some perceptual intelligibility assessments are constructed in such a way that listeners' variables are managed (e.g. using a large set of test items combined with random selection, including non-existing words and the using syntactically and grammatically correct sentences conveying no meaningful message), resulting in an acceptable level of reliability. For the Assessment of Intelligibility of Dysarthric Speech (3) Pearson's Correlation Coefficients for the inter-rater and intra-rater

reliability varied from r:0.87 to r:0.99. Also for the Dutch Intelligibility Assessment (4), which contains a large number of test sets and in which non-existing words are included, strong inter-rater and intra-rater reliability levels were found (Intraclass Correlation Coefficient (ICC) -intra-rater: 0.93; ICC-inter-rater:0.91). Another approach to obtain a reliable quantitative analysis of speech is to apply speech technology. Previous attempts to develop an automated intelligibility assessment relied on automatic speech recognition (ASR) systems that were trained to recognize speech of persons without known impairments (5,6). The acoustic models embedded in such systems are undoubtedly useful in case of normal speech sounds, substitutions, omissions and additions. However, dysarthric speech sounds often exhibit distortions and are thus far off the samples that were used during model training. This means that the models have to make strong extrapolations. Our hope is that dysarthric speech sounds can still be properly characterized in a space of articulatory features and that such a characterization offers a better basis for providing the clinician with information that is directly related to speech therapy.
The aim of the SPACE-project (Speech Algorithms for Clinical and Educational applications) is to develop a speech technology based clinical assessment tool that provides a reliable quantitative (degree of intelligibility) and qualitative (articulation errors) analysis of speech. This paper presents the preliminary results of intelligibility measures of dysarthric speech performed by 4 different speech processing systems.

METHODS

Speech samples

Digital audio-records were made of 60 dysarthric speakers by means of a Mini-disc (Sony). The subjects were instructed to read a randomly selected test set of the Dutch Intelligibility Assessment (DIA). The DIA assesses intelligibility at phoneme level. Each test set contains 50 consonant-vowel-consonant words (existing or non-existing but well pronounceable words). Each test word is constructed as a randomly selected target phoneme embedded in a fixed frame consisting of 2 non varying phonemes. As illustrated in Figure 1, a different frame is provided for each test item (1: .op, 2: .uis). Each Dutch phoneme appears at least once in each possible position (consonants: initial and final; vowels and diphthongs: medial). In the initial and final position an omission may also occur.

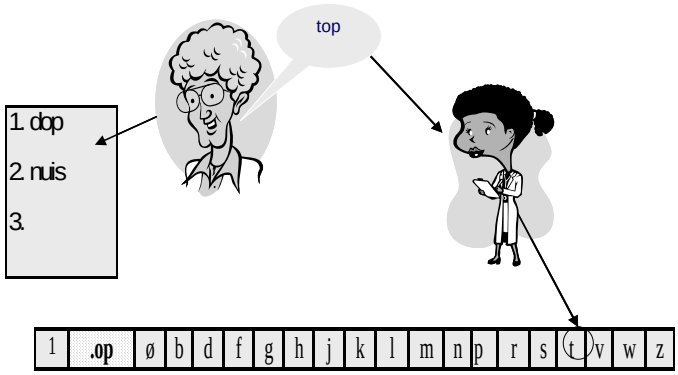

Figure 1: Graphic representation of the DIA.

Perceptual judgements and intelligibility scores

The audio recordings were judged by an experienced speech-language-pathologist (SLP) using the score sheets of the DIA. For each test item the fixed frame is presented, followed by all possible target phonemes. The SLP had to indicate the perceived phoneme. This procedure is illustrated in Figure 1. The intelligibility score is calculated as the percentage of correctly identified phonemes.

Speech processing systems (SPS): acoustic features (ACF) versus articulatory features(ARF)

Systems 1 (WAR-ACF) and 3 (CS-ACF) (7) are supplied with standard acoustic features that are used in almost all ASR-systems, namely the Mel-Frequency-Ceptral Coefficients (8). Both systems use the same (about 1000) statistical acoustic models, more precisely hidden Markov models (HMMs). Each model represent the acoustics of a phoneme when appearing in conjunction with a particular preceding and succeeding phoneme. By means of these so-called triphone models the ASR-system is able to take co-articulation phenomena into account. The models are trained on the read speech parts of the Spoken Dutch Corpus (Corpus Gesproken Nederlands) (9).

Systems 2 (WAR-ARF) and 4 (CF-ARF) (10,11) are supplied with articulatory features that were derived from the acoustic features by means of neural networks. These networks extract 25 binary articulatory features concerning voicing, vowel height, manner of articulation, place of articulation etc. The neural networks were trained on the same speech data as the triphone models. By means of 40 models describing the 40 phonemes in terms of their phonetic atoms (1 or 2 phoneme components) and their articulatory characterization, articulatory scores can easily be converted to phoneme scores, as needed by system 2. In system 4, no such conversion is required.

Speech processing systems (SPS): speech recognizer (WAR) versus speech aligner (CS)

Systems 1 (WAR-ACF) and 2 (WAR-ARF) are automatic word recognizers. For each word read by the speaker, the speech processing system uses a lexicon of all the words that can be constructed from the fixed frame (combination of two phonemes) by supplementing it with one of all the possible target phonemes. Thus, this method is in conformity with the method used for the perceptual judgments. For each utterance the speech recognizer calculates the total Log Likelihood score for each possible word in the lexicon and it selects the word with the highest score as the 'perceived' word. The degree of intelligibility is expressed as the Word Accuracy Rate (WAR) which is the percentage of correctly identified words. By means of a linear regression model, this WAR is finally converted to an 'objective' intelligibility score that is presumed to agree with the perceptual intelligibility score.

Systems 3 (CS-ACF) and 4 (CS-ARF) are speech aligners. For each word read by the speaker, the speech processing system knows what the corresponding word and its canonical (= normal or expected) phonetic transcription is. The system then segments the speech utterance in time intervals which it believes to represent the acoustic realizations of the phonemes (system 3) or phoneme components (system 4). On the basis of the segmentations computed for all utterances of a speaker, a mean confidence score (CS) per phoneme (system 3) or per articulatory feature (system 3) is computed. In the case of system 3, the confidence score of a phoneme is an estimate of the mean posterior probability of this phoneme in the speech frames observed during the different realizations of this phoneme. In the case of system 4, the confidence score of an articulatory feature is the mean posterior probability of this feature in the speech frames belonging to phoneme components possessing this feature (meaning that the feature is true). By means of a linear regression model, the confidence scores are finally converted into one 'objective' intelligibility score that is presumed to agree with the perceptual intelligibility score. The regression model first selects the most important 10 scores from the full score set and then computes the regression coefficients in the subspace of these 10 scores.

Comparison of perceptual and objective intelligibility scores

Since the number of samples is rather limited Pearson correlation coefficients (r) between the objective and the perceptual intelligibility scores are estimated by means of 5-fold cross

validation experiments. This means that the data were divided in 5 separate sets. The linear regression model was each time trained by 4 of the 5 data sets and applied on the recordings of the other data set, acting as a test set. This was repeated 5 times, until each data set was selected once as the test set. The final correlation was obtained as an average of the five Pearson correlations between the computed and the perceptual intelligibility scores.

RESULTS

The correlations for systems 1 (WAR-ACF) and 2 (WAR-ARF) were respectively found to be moderate (r:0.56) and low (r:0.33). However, the alignment-based systems resulted in much higher correlations (system 3: r:0.72; system 4: r:0.72). The correlations found for system 3 (CS-ACF) and system 4 (CS-ARF) are illustrated in figures 2 and 3. The results are summarized in Table 1.

Table 1: Correlations between the computed and perceptual scores for the 4 SPSs.

Speech recognizers		Speech aligners	
WAR-ACF	WAR-ARF	CS-ACF	CS-ARF
r:0.56	r:0.33	r:0.72	r:0.72

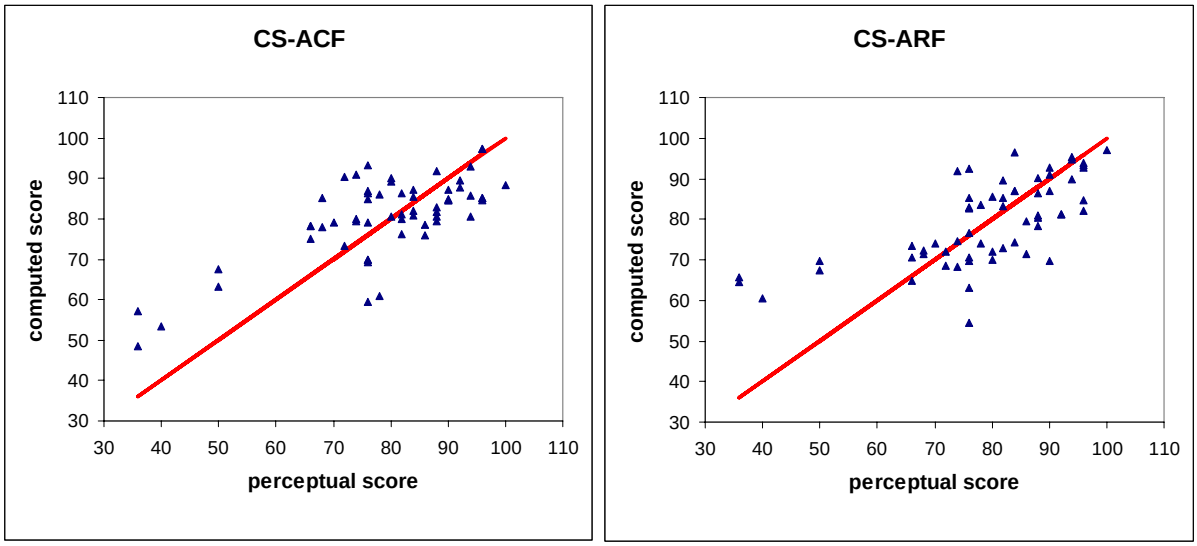

Figure 2: computed versus perceptual scores for systems 3 (CS-ACF) and 4 (CS-ARF)

DISCUSSION

The results show that in case of dysarthric speech alignment-based systems provide more reliable intelligibility scores than recognition-based systems. This is owed to the fact that the log likelihood based decision model of an ASR system is not a good model for the human decision model, and thus, that the objective and perceptual WAR are not directly comparable. The alignment systems merely produce a vector of confidence scores, representing deviations from the normative pronunciation. By means of a linear regression model, one can then try to estimate how important the different deviations are in the human decision process. The current results are encouraging but cannot yet compete with the inter-rater and intra-rater agreements found for several perceptual intelligibility assessments. There is no significant difference between the reliability of the acoustic feature based and the articulatory feature based aligner. In a later stage, we hope that the ARF system can provide the clinician with relevant information concerning the patients articulation and concerning distortions in the patient's pronunciations. However, further refinements of the ARF system are needed for this.

ACKNOWLEDGEMENT

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) Yorkston, K.M., Beukelman, D.R., Strand, E.A. & Bell, K.R. (1999). Managment of motor speech disorders in children and adults. (2nd ed.). Ausin, TX: Pro-ed.
2) Yorkston, K.M., Strand, E.A., & Kennedy, M.R.T. (1996). Comprehensibility of dysarthric speech: implications for assessment and treatment planning. American Journal of Speech-Language-Pathology, 5, 55-66.
3) Yorkston, K.M. & Beukelman, D.R. (1981). Assessment of Intelligibility of Dysarthric Speech. Austin, TX: C.C. Publishers, Inc.
4) De Bodt, M., Guns, C. & Van Nuffelen, G. (2006). NSVO: Nederlandstalig Spraakverstanbaarheidsonderzoek. Herentals: Vlaamse Vereniging voor Logopedisten.
5) Carmichael, J. & Green, Ph. (2004). Revisiting Dysarthria Assessment Intelligibility Metrics. Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP).
6) Schuster, M., Haderlein, T., Nöth, E., Lohscheller,J., Eysholdt,U. & Rosanowski,F. (2006). Intelligibility of laryngectomees' substitute speech: automatic speech Recognition and subjective rating. European Archives of Otorhinolaryngology, 263, 188-193.
7) Duchateau, J. HMM based acoustic modeling in large vocabulary speech recognition. Ph.D. dissertation, Katholieke Universiteit Leuven, November 1998, available from http://www.esat.kuleuven.be/psi/spraak
8) Davis & Mermelstein (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. On ASSP, 28, 357-366.
9) Oostdijk, N., Goedertier, F., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, H.& Baayen, H.(2002). "Experiences from the Spoken Dutch Corpus", Proceedings LREC-2002 (Las Palmas), 340-347.
10) Stouten, F. & Martens, J.P. (2005). "On the use of phonological features for pronunciation scoring", Proceedings ICASSP-2006 (Toulouse), 329-332.
11) Stouten, F. & Martens, J.P. (2006). "Speech recognition with phonological features: some issues to attend", Proceedings Interspeech-2006, 357-360.