ACOUSTIC-PHONETIC DECODING FOR SPEECH INTELLIGIBILITY EVALUATION IN THE CONTEXT OF HEAD AND NECK CANCERS

Corinne Fredouille[1], Alain Ghio[2], Imed Laaridh[3], Muriel Lalain[2], Virginie Woisard[4]

[1]Avignon Université, LIA, Avignon, France ; [2]Aix-Marseille Univ, CNRS UMR 7309, LPL, Aix-en-Provence, France ; [3]Toulouse University, CNRS UMR 5505 IRIT, Toulouse, France ; [4]Service ORL, CHU Larrey, Toulouse, France
corinne.fredouille@univ-avignon.fr,alain.ghio@univ-avignon.fr

ABSTRACT

In addition to health problems, Head and Neck Cancers (HNC) can cause serious speech disorders that can lead to partial or complete loss of speech intelligibility in some patients. The clinician's evaluation of the intelligibility level before or after surgical treatment and / or during the rehabilitation phase is an important part of the clinical assessment. Perceptive assessment is the most widely used method in clinical practice to assess the level of intelligibility of a patient despite the limitations associated with it such as subjectivity and moderate reproducibility. In this paper, we propose to overcome these limitations by associating a specific task of speech production based on pseudo-words with an automatic speech processing system, both oriented towards acoustic-phonetic decoding. Compared to human perception, the automatic system reaches very high correlation rates and promising results when applied to a French speech corpus including 41 healthy speakers and 85 patients suffering from HNC.

Keywords: acoustic-phonetic decoding, speech disorders, Head and Neck cancers, speech intelligibility, automatic speech processing

1. INTRODUCTION

We define the intelligibility by "the degree to which the speaker's intended message is recovered by the listener" [8]. More precisely, we consider that the intelligibility of a speaker corresponds to the performance by a listener to recognize the words and / or the sounds of the speech produced by the speaker. Generally, intelligibility tests are performed with sentences or words extracted from a restricted list of items. The limitation of this type of test is the ability of listeners to restore distorted sequences. This effect is the stronger as the listeners can have knowledge of the words used in the test, that these words are often unambiguous and therefore highly predictable. This is usually the case for speech therapists who can make an intensive use of such lists of items that they end up memorizing them. One example in French is the BECD [3], where the intelligibility test is based on a list of only 50 words. The bias linked to the knowledge of the closed set of items and, therefore, to the strong influence of top-down perceptive mechanisms is an overvalued intelligibility score because the phonemic restoration [17] hides the distortions of speech production. The solution proposed in this paper is to use pseudo-words in a very large quantity in order to neutralize the learning and restoration effects by the listeners. In the end, listeners are confronted with an acoustic-phonetic decoding task referred as DAP (*Décodage Acoustico-Phonétique*, i.e.: Acoustic-Phonetic Decoding) followed by a written transcription. The second proposal is to apply an automatic acoustic-phonetic decoder, specifically derived from existing automatic speech processing, which can be considered as a "robotic" listener. Indeed, if results similar to a set of human listeners are obtained, the automatic system will have the advantage to provide reproducible and deterministic behavior contrary to human perception.

2. MATERIAL AND METHOD

2.1. Construction and principles of the Acoustic-Phonetic Decoding test

For each speaker, a list of 52 items is randomly selected from a dictionary of 89346 pseudo-words. All the lists contain the same number of consonants and vowels but with different combinations, which makes the lists equivalent but different. The pseudo-words are constructed with the forms $C(C)_1V_1C(C)_2V_2$ which is conform to French phonotactic. V_i is a French Vowel, $C(C)_i$ is an isolated consonant C or a consonant group CC. These elements are selected from a list of 18 consonants and 16 consonant group of French; 8 different vowels can be selected. Examples of pseudo-words are *stoumo*, *vurtant*, *muja*, *leba*, *ranto* ...

2.2. Corpus and perceptual test

The current study is based on the French HNC speech corpus C2SI [2]. This corpus includes patients suffering from oral cavity or oropharyngeal cancer and healthy speakers. People were asked to record different speech production tasks like sustained /a/, read speech, picture description, spontaneous speech, and isolated pseudo-words. All the patients have undergone a cancer treatment consisting in surgery and/or radiotherapy and/or chemotherapy.

In this study, 85 patients and 41 healthy speakers were recorded and produced the 52 required pseudo-words. All the speech signals were segmented and the stimuli were played back randomly to 40 French native naive listeners who transcribed orthographically what they can hear. During this perceptual test of intelligibility using Perceval Station [1], each stimulus was transcribed by 3 different listeners. Once the orthographic transcriptions were collected, the objective was to extract a phonemic form because the passage through the spelling was only an intermediate step to access a phonetic representation. The orthographic transcriptions were phonetized by the LIAPHON algorithm [4] and they were compared to the expected phonetic forms of the pseudo-words. Traditionally, for ease of processing, the result is binary: correct or incorrect. To go beyond this basic evaluation, an analog result is provided here by proposing a kind of distance to the target.

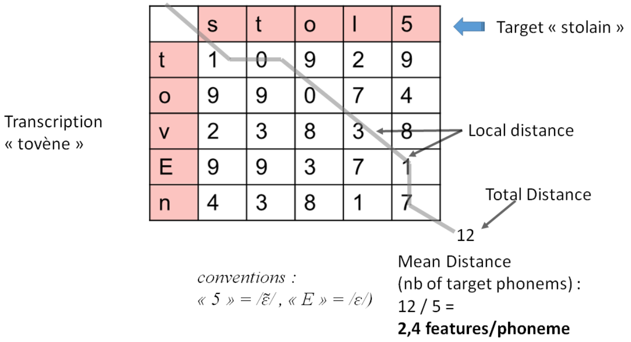
2.3. Acoustic-Phonetic Decoding and Distance measurement

To compare the perceived/transcribed form vs the expected item, a Wagner-Fischer algorithm is used that integrates the phenomena of insertion, elision and substitution of units (Figure 1).

Here, the calculation of Levenshtein distance is not based on orthographic units but on phonemes and it is possible to introduce subtle nuances because, for example, we can consider that a confusion between /a/ and /i/ does not have the same weight as between /a/ and /ã/ [7]. In a primary step, we have chosen to set the local distance by counting the number of different phonetic features between two units.

At the end, a cumulative distance between the transcription and the target is produced, which is finally normalized by the number of phonemes present in the target sequence. For instance, a final score of "2" means that there is on average a difference of 2 phonological features by phoneme between the target and the perceived form. Thus, this approach permits to obtain a score per speaker by averaging the score on the 52 pseudo-words produced by the speaker and transcribed by the listeners. The higher the score, the worse the intelligibility. This score will be denoted "Perceptual DAP-based intelligibility score" in the rest of the paper.

Figure 1: Comparison of 2 phoneme sequences by the Wagner-Fischer algorithm



	s	t	o	l	5
t	1	0	9	2	9
o	9	9	0	7	4
v	2	3	8	3	8
E	9	9	3	7	1
n	4	3	8	1	7

3. AUTOMATIC DAP-BASED TRANSCRIPTION FOR SPEECH INTELLIGIBILITY EVALUATION

3.1. Automatic DAP-based transcription

If different approaches based on automatic speech processing have been proposed in the literature do deal with the evaluation of the speech intelligibility in the context of speech disorders, [15, 5, 9, 12, 14, 10, 11, 16], they do not respond to our very specific transcription problem. The Automatic Speech Recognition (ASR) is the most suitable automatic processing to provide speech transcriptions, especially considering the great progress made over the last 10 years [13]. Nevertheless, it is well known that the ASR systems perform very well when acoustic and language models are used together in the recognition process. Here, the speech transcription is concerned with isolated pseudo-words, which totally invalidates the use of language models. Similarly, the construction methodology of the set of pseudo-words leads to a high level of confusion in the acoustic domain (*bravan* vs *brava*), possibly augmented by the speech impairment. This typical context invalidates also the application of classical automatic isolated word recognition systems. Therefore, the automatic transcription task can be simply reduced in search of the sequence of phonemes as well as their frontiers present in the given speech signal without any information about the sequence itself and the phoneme segmentation. Depending on the quality of the speech signal, this acoustic-phonetic decoding task can be far from trivial and can lead to different kinds of decoding errors like phoneme substitutions, deletions, or insertions as well as phoneme segmentation errors. To limit errors in the specific context of impaired speech concerned by this work, it is proposed to consider the automatic task of acoustic-phonetic decoding differently and to associate it with two steps as reported below.

First step : a text-constrained alignment. Given the production of a pseudo-word by a speaker, the automatic forced alignment consists in providing the temporal segmentation of the known phoneme sequence present in the speech signal. By taking as input the target pseudo-word, its sequence of phonemes and the speech signal produced by the speaker, the automatic processing is based on a decoding of the speech signal, involving the Viterbi algorithm and statistical Hidden Markov Models (HMM). Here, each phoneme is represented by a three-state context-independent HMM, estimated using the maximum likelihood estimate based on approximately 200 hours French radio recordings [6]. A Maximum A Posteriori (MAP) adaptation at 3 iterations is performed in order to create speaker-dependent models. Speech signal parameterization relies on 12 Perceptual Linear Prediction coefficients plus the energy, plus their delta and delta-delta coefficients.

Second step : a semi-constrained acoustic-phonetic decoding.

Given the phoneme segmentation obtained in the first step, the goal of this second step is to reconsider the phoneme labels and to search for the most appropriate ones among a set of 36 French phones, keeping the segment frontiers fixed (denoted as semi-constrained acoustic-phonetic decoding). In this manner, each speech segment available in the phoneme segmentation is confronted with the 36 three-state context-independent HMM implied in the previous step. The comparison of the log-likelihood measures computed between a given speech segment and these 37 acoustic models leads to a phoneme ranking. Considering all the speech segments associated with the production of a pseudo-words, a new sequence of phones is then provided.

3.2. Automatic speech intelligibility evaluation

Given the speech production of a pseudo-word, the phoneme sequence resulting from the automatic semi-constrained acoustic phonetic decoding is compared with the expected phoneme sequence composing the pseudo-word. As for the perceptual evaluation based on the human listeners, the normalized cumulative distance, described in section 2.3 between the automatic transcription and the target phoneme sequence can be computed. Still here, this normalized cumulative distance is considered as a measure of the speaker intelligibility : the higher the automatic score, the worse the intelligibility.

4. RESULTS AND DISCUSSIONS

The purpose of these experiments is to evaluate the effectiveness of the automatic DAP-based transcription for the speech intelligibility evaluation compared with the perceptual evaluation. In this way, the normalized cumulative distances were computed for the 52 pseudo-words produced by the 85 patients and the 41 healthy speakers included in the HNC corpus (6478 distances). They were then averaged to provide an automatic score of intelligibility per speaker (126 scores). Next subsections analyze the relevance of these automatic scores. The scatter plot and Linear Regression (LR) lines, including equation and determination coefficient R^2, are provided to support this analysis. For the perceptual evaluation, it is important to remember that 3 different normalized cumulative distances were computed per pseudo-word (one per listener) against only 1 distance for the automatic evaluation. Depending on the required analysis, the mean, minimum, maximum or median values could be used for the perceptual scores.

Automatic vs. perceptual DAP-based scores of intelligibility

In figure 2, scores of intelligibility - averaged normalized cumulative distances over the 52 pseudo-words per speaker (section 2.3) - from the automatic and perceptual DAP-based approach (the mean value is used here) are compared. The observation of the scatter plot shows a quite relevant correlation between both sets of scores, with a determination coefficient R^2=0.54, and, therefore, a correlation rate r=0.735. It is interesting to notice that LR lines are extremely close, illustrating the behavior of up to 50% of the automatic scores (R^2=0.54) and a factor of either 0.6 (origin-forced LR) or 0.7 (free LR) between acoustic and perceptual DAP-based scores. From these results, the DAP-based automatic evaluation tends to be more "severe" than the perceptual evaluation (yielding higher scores), but also more confused. Neverthe-

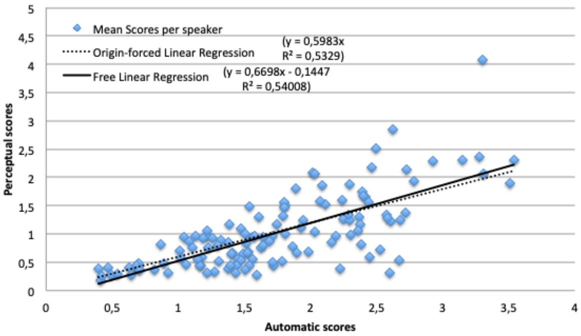


Figure 2: Comparison between the automatic and perceptual based-DAP evaluation. The scatter plot is coupled with two kinds of linear regressions (LR) : forced-origin LR (dotted line) and free LR (plain line), associated with their equation and determination coefficient R^2.

less, a preliminary analysis of automatic scores obtained at the individual pseudo-words, provided in table 1, shows that for 20% of pseudo-words, the automatic system reaches the minimum scores compared with the 3 perceptual scores of the listeners - the automatic system can be considered as the best "listener" compared with the 3 human ones. For 50% of pseudo-words, the automatic scores are less than the maximum of the 3 perceptual scores of the listeners for the pseudo-words concerned - the automatic system performs better than the "worst" listener. Finally, a difference value between the automatic and maximum perceptual scores over the 3 listeners less than 0.5 is considered, which concerns 64% of pseudo-words. This represents a confusion of 2 or less acoustic traits per pseudo-word carried out by the system, which could be rather acceptable according to the application context.

DAP-based evaluation vs human experts

Here, both automatic and perceptual DAP-based evaluation (scores per speaker) are compared with a more global assessment of the degree of intelligibility as well as of of speech disorder severity, carried out by an expert jury composed of 6 clinicians and speech therapists. This assessment was made on the task of the picture description (also available in the C2SI corpus[2]), following a 0-10 point scale (0 standing for a low intelligibility and severity). An average of 7 for severity and 8.3 for intelligibility is reached for the set of 126 speakers.

It is observed correlation rates of -0.84 vs -0.71 between the degree of intelligibility and DAP-based perceptual vs automatic scores and correlation rates of -0.85 vs -0.76 between the degree of speech disorder severity and DAP-based perceptual vs automatic scores. These results raise several questions : (1) is the automatic system more sensitive to severity than human perception, which would further disrupt its acoustic-phonetic decoding? (2) no difference is observed for the human perception between intelligibility and severity in terms of correlation rates whereas more variation is underlined in the global assessment between both criteria in [2]. Could it mean that we have reached the ceiling that human perception based on acoustico-phonetic decoding can achieve in terms of accuracy of the evaluation ?

Table 1: Preliminary analysis of the set of pseudo-words-based automatic scores compared with the set of perceptual scores. Minimum (*min*) and maximum (*max*) statistics are applied considering the perceptual scores of the 3 listeners.

Statistics type	values
Nb pseudo-words-based scores	6477
Nb auto. scores <= min(perceptual scores)	1280
Nb automatic scores <= max(perceptual scores)	3242
diff(auto. scores, max(percept. scores) <= 0.5	4161

5. CONCLUSIONS AND PERSPECTIVES

The goal of this paper is to present an original approach for the clinical evaluation of speech intelligibility, based on a typical generation of pseudo-words for the speech production and their acoustic-phonetic decoding by an automatic speech processing for the evaluation step. Applied to a corpus of patients suffering from Head and Neck Cancers, this original approach could be easily generalized to the evaluation of any speech disorders like dysarthria for instance. Experimental results underline the potentiality of such an automatic approach (for 50% of pseudo-words considered in this study, the automatic system performs better than the "worst" human listener), for which the objectivity, the reproducibility as well as the deterministic behavior are undeniable faced to the human perception. A thorough analysis of the behavior of the automatic system according to the different phonetic classes present in the list of pseudo-words will be the next step of our future work. Indeed, it will be very relevant to observe whether certain phonetic contexts facilitate or not the automatic or perceptual measurement of intelligibility according to the degree of speech disorders observed in the patient. This analysis should make it possible to answer both questions raised when comparing the evaluation based on DAP with the overall assessment of the degree of intelligibility and severity.

6. ACKNOWLEDGEMENTS

This work has been carried out thanks to the French National Cancer Institute (INCA) in 2014 C2SI project (Grant n^{o}2014-1), and to the French National Research Agency in 2018 RUGBI project untitled "Improving the measurement of intelligibility of pathological production disorders impaired speech" (Grant n^{o}ANR-18-CE45-0008).

7. REFERENCES

[1] André, C., Ghio, A., Cavé, T. B., Christian 2003. Evaluation of a phone-based anomaly detection approach for dysarthric speech. *Proceedings of Intl Congress of Phonetic Sciences (ICPhS'03)* Barcelone, Spain.

[2] Astesano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pont, O., Pouchoulin, G., Michele, P., Robert, D., Sicard, E., Woisard, V. may 2018. Carcinologic Speech Severity Index Project: A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer. *Language Resources and Evaluation Conference (LREC), Miyazaki, Japon* http://www.elra.info. European Language Resources Association (ELRA).

[3] Auzou, P., Rolland-Monnoury, V. 2006. *Batterie d'évaluation clinique de la dysarthrie*. Édition Ortho.

[4] Bechet, F. 2001. Liaphon : Un système complet de phonétisation de textes. *Traitement Automatique des Langues* 42, 47–67.

[5] Christensen, H., Cunningham, S., Fox, C., Green, P., Hain, T. 2012. A comparative study of adaptive, automatic recognition of disordered speech. *Proceedings of Interspeech'12* Portland, USA.

[6] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G. September 2005. ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *Proceedings of Interspeech'05* 1149–1152.

[7] Ghio, A., Rossi, M. 1995. Parallel distributed processes for speaker independent acoustic- phonetic decoding. *International Congress of Phonetic Sciences (ICPhS)* volume 4 Stockholm, Sweden. 272–275.

[8] Kent, R. D., Weismer, G., Kent, J.-F., Rosenbek, J. 1989. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders* 54, 482–499.

[9] Khan, T., Westin, J., Dougherty, M. 2014. Classification of speech intelligibility in parkinson's disease. *Biocybernetics and Biomedical Engineering* 34(1), 35–45.

[10] Kim, M., Cao, B., An, K., Wang, J. 2018. Dysarthric speech recognition using convolutional lstm neural network. *Proceedings of Interspeech'18* Hyderabad, India.

[11] Laaridh, I., Fredouille, C., Ghio, A., Lalain, M., Woisard, V. 2018. Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers. *Proceedings of Interspeech'18* Hyderabad, India. 2943–2947.

[12] Laaridh, I., Fredouille, C., Meunier, C. May 2015. Automatic detection of phone-based anomalies in dysarthric speech. *ACM Transactions on accessible computing* 6(3), 9:1–9:24.

[13] Maier, A. K., Schuster, M., Batliner, A., Nöth, E., Nkenke, E. 2007. Automatic scoring of the intelligibility in patients with cancer of the oral cavity. 1206–1209.

[14] Martínez, D., Lleida, E., Green, P., Christensen, H., Ortega, A., Miguel, A. 2015. Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing (TACCESS)* 6(3), 10.

[15] Middag, C., Martens, J.-P., Van Nuffelen, G., De Bodt, M. 2009. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing* 2009(1), 1–9.

[16] Vasquez-Correa, J., Orozco-Arroyave, J., Bocklet, T., Noth, E. 2018. Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. *Journal of Communication Disorders* 76, 21–36.

[17] Warren, M. R. 1970. Perceptual restoration of missing speech sounds. *Science* 167"3917", 392–395.