INTERSPEECH 2014

The Goodness of Pronunciation algorithm applied to disordered speech

Thomas Pellegrini[1], *Lionel Fontan*[1], *Julie Mauclair*[1,2], *Jérôme Farinas*[1], *Marina Robert*[3]

[1]Université de Toulouse; UPS; IRIT; Toulouse, France
[2]Université Paris Descartes, Paris, France
[3]Université Paris Ouest, Paris, France

pellegri,fontan,mauclair,jfarinas@irit.fr

Abstract

In this paper, we report on a study with the aim of automatically detecting phoneme-level mispronunciations in 32 French speakers suffering from unilateral facial palsy at four different clinical severity grades. We sought to determine if the Goodness of Pronunciation (GOP) algorithm, which is commonly used in Computer-Assisted Language Learning systems to detect learners' *individual errors*, could also detect segmental deviances in disordered speech. For this purpose, speech read by the 32 speakers was aligned and GOP scores were computed for each phone realization. The highest scores, which indicate large dissimilarities with standard phone realizations, were obtained for the most severely impaired speakers. The corresponding speech subset was manually transcribed at phone-level. 8.3% of the phones differed from standard pronunciations extracted from our lexicon. The GOP technique allowed to detect 70.2% of mispronunciations with an equal rate of about 30% of false rejections and false acceptances. The phone substitutions detected by the algorithm confirmed that some of the speakers have difficulties to produce bilabial plosives, and showed that other sounds such as sibilants are prone to mispronunciation. Another interesting finding was the fact that speakers diagnosed with a same pathology grade do not necessarily share the same pronunciation issues.

Index Terms: pronunciation automatic assessment, Goodness of Pronunciation, disordered speech

1. Introduction

Unilateral facial palsy (UFP) can result from a variety of causes, such as trauma, infection or tumors [1]. Among its numerous consequences on patients' lives, UFP often cause articulatory disorders than can greatly impact on communication ability [2]. However, the assessment of the severity of impairment generally relies on physical criteria only, such as in the House-Brackmann scale [3, 4]. This clinical tool evaluates the degree to which patients can activate mouth, eyelids and forehead muscles when executing voluntary or involuntary movements, and leads to a score ranging from grade I (normal facial activity) to grade VI (total palsy).

For high pathology grades, the inability to control the lips hinders a proper control of the air flow. Phonemes most impacted are consonants: bilabials /p, b, m/ may lose their burst phase, labiodentals /f, v/ and fricatives /s, S/[1] are also impacted due to an unilateral stretching/ closing of the lips [6]. A qualitative study showed that the most affected consonants are /p/ and /f/ [7]. To a lesser extent, vowels may also be affected, in particular vowels that imply a certain control of the lips' movements such as /e, i, o, u, y/ and nasal rounded vowels /o~, e~/. If studies reported a clear correlation between the severity of impairment and the articulatory disorders, a large variability in the articulation abilities of impaired speakers was also observed, even among speakers sharing a same palsy grade [6].

Having an automatic tool to assess pronunciation at phone-level would allow to easily gather individualized information about each patient. Assessment of speech abilities is indeed very time-consuming, which does not necessarily fit clinical means for patients' evaluation. Such tools are widely used in Computer-Assisted Language Learning (CALL) systems, with early works reported in the 1990s [8]. CALL systems use Automatic Speech Recognition (ASR) techniques to assess non-native pronunciation both at suprasegmental and segmental levels, two domains respectively referred to as *overall pronunciation assessment* and *individual error detection* [9]. Within the latter scope, several scoring approaches can be used to detect phoneme mispronunciations. They range from the analysis of raw recognition scores [10], likelihood ratios such as native-likeness and Goodness of Pronunciation (GOP), to the definition of scores derived from classification methods such as linear discriminant analysis and alike [11]. Contrary to native-likeness scores that imply the use of non-native acoustic models, the Goodness of Pronunciation (GOP) algorithm is solely based on the comparison of speakers' realizations with native phone models. It calculates a likelihood ratio indicating the degree to which a phone may be the realization of a specific phoneme of the target language [12, 13]. In other words, GOP scores give an idea of how distinguishable is a specific phone realization compared to other phones, and can thus be thought of as intelligibility indexes [14]. As a consequence, the relevance of GOP scores for speakers' evaluation may not be limited to the framework of foreign language learning, but possibly applies to any kind of articulatory deviances, such as in motor speech disorders.

Studies found in the literature mostly focus on ASR system performance and limits when processing disordered speech, especially concerning dysarthric speech [15, 16, 17, 18]. In [19], ASR word accuracy was correlated with subjective speech intelligibility for children with cleft lip and palate. However, to the best of our knowledge, the present study is a first attempt to use both ASR and CALL techniques to assess the pronunciation skills of impaired speakers — in the particular case of patients suffering from unilateral facial palsy. The two main questions addressed in this study are: Can the GOP algorithm be used to identify and characterize individual mispronunciations in the context of peripheral paralysis impairments? Does the GOP scores correlate with clinical impairment grades?

[1] We will use the SAMPA phonetic alphabet [5] throughout the paper

The paper is organized as follows. First, an overview of the GOP algorithm is given. Sections 3 and 4 describe the methodology and the speech corpus used in this work, followed by a listening analysis of the corpus. Statistics on manual transcriptions at phone level are then described. Finally, GOP experiments are reported and discussed in Section 7.

2. The GOP algorithm

To compute GOP scores on a given utterance, two phases are needed: 1) a free speech recognition phase and 2) a forced alignment phase. Without giving any information to the ASR system about the target sentence, the free speech recognition phase determines the most likely phone sequence matching the audio input (*i.e.* the output is that of a free phone loop recognizer). On the contrary, the forced alignment phase implies to provide the ASR system with the orthographic transcription of the input sentence. It then consists of forcing the system to align the speech signal with the expected phone sequence.

For each phone realization aligned to the speech signal, a GOP score is calculated by taking the absolute value of the difference between the log-likelihood of the forced alignment phase and the one of the free recognition phase. When the expected and the freely recognized phones are the same, the GOP score is zero. Otherwise, the larger the GOP score, the greater the probability of a mispronunciation. In order to decide whether a phone was mispronounced ("rejected") or not ("accepted"), phone-specific thresholds need to be defined. In this work, we used the baseline implementation of the GOP algorithm, described in [12, 13].

3. Methodology

First, a preliminary auditory analysis of the speech corpus was conducted in order to identify pronunciation trends that could discriminate between speakers suffering from several palsy severity grades. Second, the GOP algorithm was run over the corpus. The forced alignments were constrained by standard pronunciations taken from a 62K French words lexicon. The aligned phone sequences were manually edited by an annotator with a solid background in phonetics and experience in transcribing speech in the context of French as a foreign language (FFL) teaching. Phones that were edited by the annotator differed from standard pronunciations and therefore were considered as mispronunciations. The resulting manual phone transcriptions were taken as *groundtruth* reference to quantify the pronunciation issues observed during the listening analysis, and also to assess the effectiveness of the GOP algorithm.

In this study, we limited this manual effort to the group of most impaired speakers group (grades V&VI UFP, *cf.* next section), since we expected more mispronunciations in this population. Furthermore, we made the assumption that all the phone sequences aligned automatically for the control group (no pathology) were correct and then taken as phone realizations that the GOP algorithm should accept. We set phone-dependent GOP score thresholds by limiting the false rejection (FR) rate below 10% on the control group. We will also report results for the operating point where false rejection and acceptance rates are equal.

4. Corpus description

The experiments have been carried out on a subset of a read speech database recorded at the *La Pitié Salpétrière* Hospital in Paris, France. This database was used in previous studies [6, 20]. It was collected from 32 French speakers suffering from UFP at five different grades according to the House and Brackmann scale, namely grades I, III, IV, V and VI. As the patients whose UFP grade had been rated as V or VI did not differ in terms of lips mobility, we regrouped them into a single group. As a result, 4 speaker groups were defined for this study. To simplify the notation in the remainder of the article, we will refer to the groups G1 (control group), G2 (grade III), G3 (grade IV) and G4 (grades V&VI). The 32 speakers are evenly distributed into these four groups, with equal proportions of male and female speakers, and a mean age of 45 years.

The patients were recorded in a soundproof booth. They read aloud 17 declarative sentences, which included all standard French consonants and semi-consonants[2]. Each sentence was constructed in order a) to include different realizations of a target consonant (i.e. alliterations) or b) to lead the speaker to produce several times a specific phonetic contrast (e.g. voicing). As an example, the sentence "Le moteur de ma moto n'a pas démarré" (*The engine of my motorbike did not start*) used an alliteration of the bilabial consonant /m/, and the sentence "Le catalogue de Paul est tombé" (*Paul's catalog fell down*) relied on the production of voiced/voiceless stop pairs (/g/ vs. /k/, /d/ vs. /t/ and /b/ vs. /p/). More details about the speech corpus can be found in [6].

5. Listening analysis

A preliminary listening analysis was conducted by two researchers familiar with pathological speech analysis. As occurred to both listeners, G3 speakers' performance seemed less impacted than that of G2 speakers. Their reading was more fluent. Although differences with the control group were found for these two groups, the deviances for the eight G4 speakers were much more important. Another general impression was that these deviances strongly depend on the speaker, even among speakers in the same group.

Perceptively, bilabial and labiodental consonants /p, f, b/, and /v/ were identified as the most impacted phonemes. Voiced phonemes (/b, v/) realizations were judged very hard to perceive, and their voiceless counterparts (/p, f/) realizations were often perceived as too breathy. A lack of control of the air flow — often referred to as *breathiness* factor in clinical tools for disordered speech evaluation — could explain this impression. Moreover, the patients' difficulties to move their lips in order to produce explosions may explain why occlusive consonants (/p, b/) were sometimes perceived as constrictive ones (/f, v/). For example, the first name *Paul* (/pOl/) was sometimes perceived as [fOl]. For some speakers /b/ realizations often sounded as [v], as in the word *bu* (/by/, *drunk*) perceived as [vy]. Concerning the fricative /v/ realizations, they have been perceived as the semi consonant /H/ in some speakers. More generally, exaggerated breathiness impacted all voiceless occlusive consonants (/p, t, k/), which were perceived with an aspiration such as in English realizations. A tendency to produce retroflex variants of phonemes /S/ and /d/ was also perceived for some speakers. This could be the result of a strategy to compensate the lack of articulation possibilities in the mouth and lips by lifting the tongue towards the post-alveolar area.

Quite unexpectedly, no peculiarities were identified for /m/ and /n/ realizations, although their place of articulation could

[2] Standard French phonological system includes bilabial (/p/, /b/, /m/), labiodental (/f/, /v/), alveodental (/t/, /d/, /n/), alveolar (/s/, /z/, /l/), palatal (/S/, /Z/), velar (/k/, /g/) and uvular (/R/) consonants, as well as the three semi-consonants /w/, /H/ and /j/

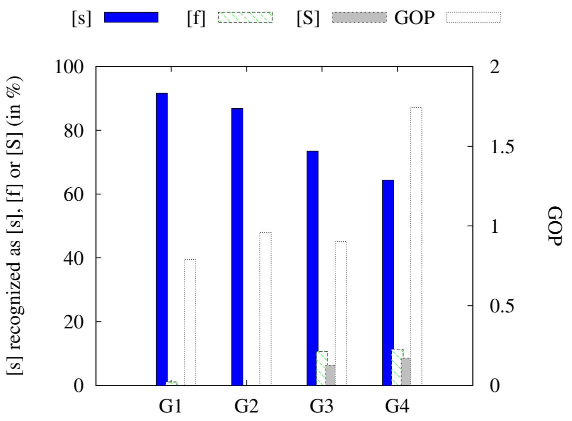

Figure 1: *GOP average scores and most frequent substitutions of the sibilant /s/, as found by the phone recognizer.*

have been thought as problematic for patients suffering from UFP. This might be explained by the fact that these two nasal consonants generate much weaker bursts than their oral counterparts, and thus that their intelligibility may be less impacted by speakers' lips hypokinesia.

6. Manual phone-level transcription

In this section, only speakers from group G4 are concerned since manual corrections were realized for this group only. Of the total 4K phones that were automatically aligned, 8.3% differ after manual corrections, with 3.6%, 2.5%, and 2.2% of substitutions, insertions, and deletions, respectively. The annotator was free to use phonetic symbols borrowed to phonological systems other than French, but we limited the present study to the French phonemic inventory. For instance, he introduced aspirated variants of /p, t, k/ phonemes that are not taken into account in the numbers reported here.

Insertions were mainly additions of schwas (1.3%). The lexicon comprises pronunciation alternates with and without schwas, whose realization is optional in French. Hence, the automatic recognizer seems to have the tendency to use pronunciations with eluded schwas when aligning. The most frequent manual editions were, in decreasing importance, deletions of voiced and unvoiced plosive closures (1.1%), substitutions of [p] by [f] (0.7%), by [b] (0.1%), by [w] and [v] (0.05%), deletions of [t] (0.6%), insertions of [Z] (0.2%) after a [d], substitutions of [b] by [v]. These results confirmed most of the observations made during the listening analysis, concerning both the impacted phonemes and the inter-speaker variability. The frequent substitutions of [p] by [f] concerned half of the speakers of G4. This observation confirms our previous study, in which the presence of a burst for /p/ was a crucial feature used in determining automatically the UFP grade [20].

7. GOP experiments

7.1. Setup

The alignment and recognition setup consists of three state left-to-right HMMs with 32 Gaussian mixture components trained on the ESTER corpus [21]. Context-independent acoustic models (39 monophones) were used since they have been found to be more suitable for CALL applications than context-dependent units [14]. This work was carried out with HTK [22].

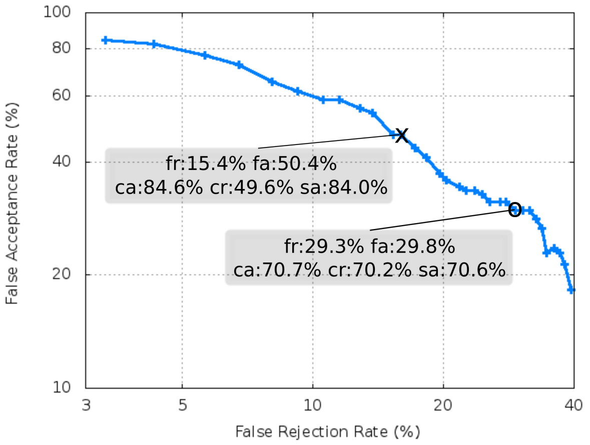

Figure 2: *DET curve for G4. Two special operating points are represented: 'x' with FR=10% on G1 data, and 'o' with equal FR and FA on G4 data.*

7.2. Inter-group results

In Table 1, we report the average and standard deviations of GOP values for each speaker group. It appeared that means and standard deviations globally increased with the impairment grade (t-tests gave p-values$<$0.01), except for group G3. A relative increase of about 34% was found between groups G1 and G4, for instance. Group G3 showed a smaller GOP mean compared to G2, which confirmed our impression that speakers from this group sounded less impacted than G2 speakers.

Figure 1 shows the most frequent substitutions of the sibilant /s/ with other consonants, as found by the phone recognizer. Average GOP scores are also given for the four groups. As one can see, confusions with [f] and [S] increase with the impairment grade, such as the average GOP for this phone, except for G3. Nevertheless, the GOP score evolution was not always that clear for other phones. An attempt to cluster the speakers according to their GOP values revealed unsuccessful, confirming our impression gained during the listening analysis that the speakers of our database, even the most impaired ones, do not share the same pronunciation issues.

Table 1: *Average and standard deviation (std) GOP values per group*

Group	Average GOP (std)
G1	1.68 (2.98)
G2	1.94 (3.12)
G3	1.72 (2.86)
G4	2.25 (3.50)

7.3. GOP algorithm accuracy

The GOP algorithm was evaluated in terms of Scoring Accuracy (SA), which is the percentage of Correct Acceptances (CA) and Correct Rejections (CR) divided by the total number of tokens (N): $SA = 100\% * (CA + CR)/N$. Correct acceptances are phones that were correctly pronounced and whose GOP scores were below a given threshold. Correct rejections are phones that were pronounced incorrectly and whose GOP scores were above a given threshold. To give an idea of performance obtained in CALL applications, SAs of about 80% obtained on Dutch non-native speech, with 50% for CA and 32% for CR, were reported on a test set [23].

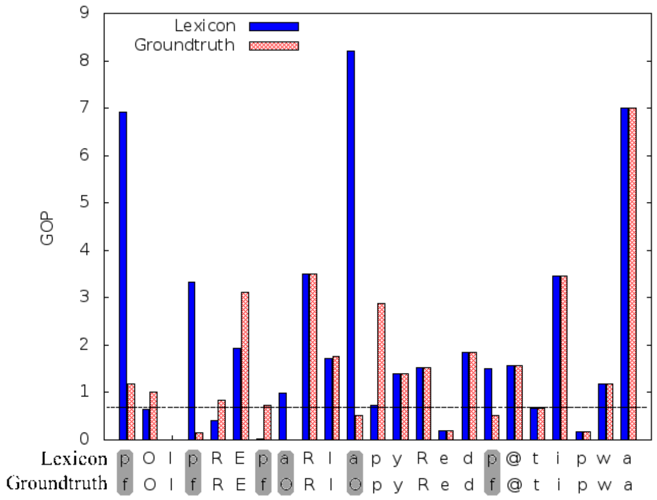

Figure 3: *Example of GOP scores for the utterance "Paul prépare la purée de petits pois". "Lexicon": phones extracted from a French lexicon, "Groundtruth": phones identified by the annotator. Horizontal dashed line: rejection threshold for phone /p/ (0.7).*

In Section 6, we reported that 8.3% of the 4K phones aligned with speech of the G4 speakers were edited by the annotator, of which 2.5% were insertions. Insertions cannot be handled by the algorithm since no GOP scores are computed for them. Hence, the algorithm should detect at most 233 mispronunciations, which correspond to 5.8% of 4K realizations.

We set phone-specific thresholds based on the GOP scores obtained on the phone realizations from the control group speakers (G1). We considered all their realizations to be correct. As a consequence, rejections are only due to misrecognitions or misalignments made by the system. Thresholds were defined by carrying out an exhaustive search for each phone, starting from a zero value with a step size of 0.1. To illustrate the detection performance, Figure 2 shows a Detection Error Tradeoff (DET) curve obtained on the G4 data. In [23], the use of a false rejection (FR) rate below 10% was justified by the fact that false rejections are more detrimental than false acceptances for a learner. With such a criterion applied on our data, we achieved SA, CA, and CR rates of 84.0%, 84.6%, and 49.6%, respectively. This operating point is indicated on the DET curve by a cross. Correct acceptance rate depends on the phones. The GOP threshold for /p/, for instance, was set to 0.7 with this configuration, and the most frequent mispronunciation, /p/ pronounced as an [f], was correctly detected in 60% of the cases. Another operating point is highlighted by a circle on the graph. It corresponds to the point with equal FR and FA rates (about 30%) obtained on the G4 test data. At this point, the system detected 70.2% mispronunciations, but the correct acceptance rate decreased to 70.7% compared to 84.6% obtained at the previous operating point.

Figure 3 illustrates an example of GOP scores obtained on the utterance "Paul prépare la purée de petits pois" (*Paul is preparing mashed peas*) for a speaker of group G4 with difficulties to produce /p/ consonants. This figure compares two sets of GOP scores computed with two different phone transcriptions given below the x-axis: one with standard pronunciations, indicated by the label "*Lexicon*", and the manual one, indicated by the label "*Groundtruth*". Six differences between the two sequences were highlighted with a Gray background color: four /p/ realizations and two /a/ realizations transcribed by [f] and [O] respectively. The horizontal dashed line of equation GOP=0.7 was added to indicate the threshold for phone [p]. Every *Lexicon* GOP score (bars in blue color) above this line gives a rejection. As can be seen, all the mispronunciations of the phoneme /p/ were correctly rejected with this threshold, but one correct pronunciation (in the middle of the utterance) was incorrectly rejected (score=0.72). Since *Groundtruth* corresponds to phones that were actually pronounced, the corresponding GOP scores are expected to be smaller than the lexicon-based ones. Indeed the GOP average values for the whole subcorpus were 2.03 and 2.25 for *Groundtruth* and *Lexicon*, respectively.

8. Conclusions

In this paper, we reported our findings from a detailed analysis of pronunciation at phone-level of speakers suffering from unilateral facial palsy at different clinical severity grades. A read speech corpus recorded from 32 French native speakers was used at this purpose. Mispronunciations were identified automatically by using the GOP algorithm originating from the CALL research area. It proved to be effective by correctly detecting 49.6% and 84.6% of mispronunciations (CR rate) and correct pronunciations, respectively, when allowing a false rejection rate of only 10% on the control group speech used to set the GOP phone-specific thresholds. CR rate increased to 70.2% with equal FR and FA rates of about 30%. The highest average GOP scores, which indicate large deviances from standard phone realizations, were obtained with speech of the most impaired speakers.

Nevertheless, average GOP scores did not strongly correlate with clinical severity grades, as measured through the House-Brackmann scale. In our opinion, this result might reflect that (a) the H&B scale has been designed to evaluate the *overall* facial mobility – i.e. not that of speech articulators only – and (b) that some patients may employ efficient compensatory strategies in order to remain intelligible despite the motor deficiencies they are suffering from. In this perspective an automatic tool such as the GOP would constitute an interesting means to easily collect relevant data pertaining to the communication ability of patients.

As a consequence, future work will be conducted in order to study further the validity of GOP measures for clinical applications, both concerning Unilateral Facial Palsies and other motor speech disorders.

9. References

[1] Ljostad, U. and Okstad, S. and Topstad, T. and Mygland, A. and Monstad, P., "Acute peripheral facial palsy in adults," *J Neurol.*, vol. 252(6), pp. 672–676, 2005.

[2] P. Gatignol and G. Lamas, *Paralysies faciales*. Solal Editions, 2004.

[3] J. House and D. Brackmann, "Facial nerve grading system," *Otolaryngology-Head and Neck Surgery*, vol. 93, pp. 146 – 147, 1985.

[4] R. Evans, M. Harries, D. Baguley, and D. Moffat, "Reliability of the House and Brackmann grading system for facial palsy," *J Laryngol Otol*, vol. 103(11), pp. 1045–1046, 1989.

[5] J. C. wells, "Sampa computer readable phonetic alphabet," in *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore, and R. Winski, Eds. Berlin and New York: Mouton de Gruyter, 1997.

[6] M. Robert, J. Mauclair, E. Lannadere, F. Tankéré, G. Lamas, and P. Gatignol, "Analyse des troubles articulatoires au cours des paralysies faciales périphériques," *LXVII Congrès de la Société Française de Phoniatrie*, 2011.

[7] A.-C. Albinhac and A. Rodier, "Analyse quantitative et qualitative des troubles d'articulation dans les paralysies faciales périphériques," Mémoire pour l'obtention du CCO, Paris, 2003.

[8] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in English pronunciation," in *Proc. ICSLP*, Kobe, 1990.

[9] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[10] B. Sevenster, G. d. Krom, and G. Bloothooft, "Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs," in *Proc. STiLL*, Marholmen, 1998, pp. 91–94.

[11] H. Strik, K. P. Truong, F. de Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection." in *Proc. INTERSPEECH*, 2007, pp. 1837–1840.

[12] S. Witt, "Use of Speech Recognition in Computer-Assisted Language Learning," PhD Thesis, University of Cambridge, Dept. of Engineering, 1999.

[13] S. Witt and S. Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning," vol. 30, pp. 95–108, 2000.

[14] T. Kawahara and N. Minematsu, *Tutorial on CALL Systems at Interspeech*, Portland, 2012.

[15] C. Coleman and L. Meyers, "Computer recognition of the speech of adults with cerebral palsy and dysarthria," *Augmentative and Alternative Communication*, vol. 7:1, pp. 34–42, 1991.

[16] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augmentative and Alternative Communication*, vol. 16:1, pp. 48–60, 2000.

[17] E. Sanders, M. Ruiter, L. Beijer, and H. Strik, "Automatic recognition of dutch dysarthric speech: A pilot study," in *Proc. ICSLP*, Denver, 2002.

[18] F. Rudzicz, "Articulatory Knowledge in the Recognition of Dysarthric Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, May 2011.

[19] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *Int J Pediatr Otorhinolaryngol*, vol. 70(10), pp. 1741–1747, Oct. 2006.

[20] J. Mauclair, L. Koenig, M. Robert, and P. Gatignol, "Burst-based features for the classification of pathological voices," in *Proc. Interspeech*, Lyon, Aot 2013, pp. 2167–2171.

[21] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Proc. Interspeech*, 2005, pp. 1149–1152.

[22] S. Young and S. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[23] S. Kanters, C. Cucchiarini, and H. Strik, "The Goodness of Pronunciation Algorithm: a Detailed Performance Study," in *SLaTE 2009 - 2009 ISCA Workshop on Speech and Language Technology in Education*, 2009, pp. 2–5.