Consonant-to-Vowel/Vowel-to-Consonant Transitions to Analyze the Speech of Cochlear Implant Users

T. Arias-Vergara[1,2,3], J.R. Orozco-Arroyave[1,2], S. Gollwitzer[3], M. Schuster[3], E. Nöth[2]

[1] Faculty of engineering. Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
[2]Pattern Recognition Lab. Friedrich-Alexander University, Erlangen-Nürnberg, Germany
[3] Department of Otorhinolaryngology, Head and Neck Surgery. Ludwig-Maximilians University, Munich, Germany
tomas.ariasvergara@lmu.de

Abstract. People with postlingual onset of deafness often present speech production problems even after hearing rehabilitation by cochlear implantation. In this paper, the speech of 20 postlingual (aged between 33 and 78 years old) and 20 healthy control (aged between 31 and 62 years old) German native speakers is analyzed considering acoustic features extracted from Consonant-to-Vowel (CV) and Vowel-to-Consonant (VC) transitions. The transitions are analyzed with reference to the manner of articulation of consonants according to 5 groups: nasals, sibilants, fricatives, voiced stops, and voiceless stops. Automatic classification between cochlear implant (CI) users and healthy speakers shows accuracies of up to 93%. Considering CV transitions, it is possible to detect specific features of altered speech of CI users. More features are to be evaluated in the future. A comprehensive evaluation of speech changes of CI users will help in the rehabilitation after deafening.

Keywords: Hearing loss, Acoustic analysis, automatic classification, Cochlear implant

1 Introduction

Hearing loss can affect speech production in both adults and children. People suffering from severe to profound deafness may experience different speech disorders such as decreased intelligibility, changes in terms of articulation, increased or decreased nasality, slower speaking rate, and decreased variability in fundamental frequency (F0) [4, 8, 10]. Furthermore, speech disorders vary depending on the age of occurrence of deafness. When hearing loss occurs after speech acquisition (postlingual onset of deafness), speech impairments are caused by the lack of sufficient and stable auditory feedback [9]. Currently, there are different treatments available for different types and degrees of hearing loss. Cochlear Implants (CI) are the most suitable devices for severe and profound deafness when hearing aids do not improve sufficiently speech perception. CI consists of an outer part, the speech processor, where acoustic information is transformed into electrical stimuli that are forwarded through the skin to the implanted part that goes into the cochlea. Due to the frequency distribution along the cochlear length,

the electric stimuli can provide frequency information. However, CI users often present altered speech production and limited understanding even after hearing rehabilitation. If the deficits of speech would be known the rehabilitation might be adequately addressed. Previous studies have analyzed speech disorders in postlingual CI users. In [3], a study was presented to evaluate hypernasality considering speech recordings of 25 postlingual CI users and 25 age-matched Healthy Controls (HC). Nasometric measures were obtained using two sentences uttered by patients and controls. The authors reported higher nasalance scores in the CI users compared to the healthy speakers. In [18], a study was presented considering speech recordings of 40 postlingual CI users and 12 HC speakers. Acoustic analysis was performed computing the fundamental frequency ($F0$) from the sustained phonation of vowel /a/. The authors reported a reduction of $F0$ in the CI users compared to the control group. In [6], a study was presented to evaluate speech deterioration in 3 postlingually deafened adults. Additionally, speech recordings of 3 HC speakers were considered for comparison. The authors reported greater $F0$ variability in the CI users compared with the control group. Furthermore, the patients showed less differentiation of place of articulation in fricative and plosive consonants.

This paper investigates the use of Consonant-to-Vowel (CV) and Vowel-to-Consonant (VC) transitions to detect speech problems in postlingual deafened CI users. Furthermore, the consonants are labeled as nasals, sibilants, fricatives, voiced stops, and voiceless stops. The reason is that the articulatory settings necessary to produce certain speech sounds may be altered in the hearing impaired people. As described in [7], after cochlear implantation, the user may notice differences between the sounds perceived and the sounds produced. If this is the case, then the CI user will move the articulators in order to produce a speech sound similar to the sound perceived. Such changes may be captured with the transitions. Previous work has considered voiced-to-voiceless/voiceless-to-voiced transitions to evaluate altered articulatory motor control in neurological diseases such as Parkinson's disease [13]. In the present study, the transitions are extracted considering different phoneme groups in order to detect speech production problems in CI users. In the neural model of speech production (DIVA model) proposed in [2], speech movements are planned considering phoneme-specific and speaker-specific mappings, which are acquired and maintained with the use of auditory feedback. With ongoing hearing loss the speech sound map can slightly change, but moreover, the sensory-motor control is decreasing as one tends to use only as much force and effort for all movements as necessary. Therewith articulation looses its precision. The use of transitions in this study is also motivated by previous findings related to speech motor control. As proposed in [14], the production of speech sound sequences is based on acoustic goals. For example, for many consonants one goal is the abrupt acoustic transition to surrounding vowels associated with a diminution of sound level. According to [14], an internal model is used by the brain to control the necessary articulatory movements to achieve different acoustic goals. Such an internal model is acquired and maintained with the use of auditory feedback. Thus, speech disorders occur when there is no sufficient speech perception. For instance, note that voiceless-stop-to-vowel transitions might be correlated with the voice onset time, which is defined as the time between the release of the oral constriction for any plosive sound and the beginning of

vocal fold vibrations [11] and has been considered in other studies to evaluate voicing contrast in CI users [7]. On the other hand, previous work suggests that fricative and sibilant production differs between CI users and HC speakers. Particularly for sibilants, these changes are produced because the spectral resolution of the CIs is lower in higher frequencies, thus, CI users shift the production of the sibilant sounds into the frequency range perceived by them [12]. We caution that it is not the aim of our study to find acoustic goals (as described in [14]), but to detect speech problems in CI users by extracting different acoustic features from the CV/VC transitions. We believe that such an approach will lead to the development of computational tools that will help to adapt hearing rehabilitation and speech therapy to the specific needs of CI users.
The rest of the paper is organized as follows: Section 2 includes details of the data and methods. Section 3 describes the experiments and results. Section 4 provides conclusions derived from this work.

2 Materials and methods

Figure 1 shows the methodology proposed in this work. First, forced alignment is performed over the speech recordings uttered by each speaker. Then, the phonemes are labeled as vowels, nasals, sibilants, fricatives, voiced stops, and voiceless stops. Then, the CV/VC transitions are extracted and assigned to their corresponding phoneme group. In the next step, acoustic features are extracted for each group of CV/VC transitions and then a Support Vector Machine (SVM) is considered to classify between CI users and healthy controls (HC). Each stage of the methodology is described in more detail in the following sections.

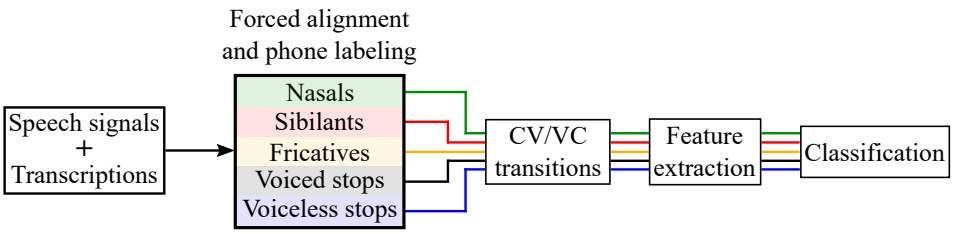

Fig. 1. Methodology implemented in this study.

2.1 Data

Standardized speech recordings of 20 postlingual deafened CI users (4 men) and 20 healthy controls (11 men) German native speakers were considered for the tests. The speech signals were captured in noise-controlled conditions at the Clinic of the Ludwig-Maximilians University in Munich, with a sampling frequency of 44.1 kHz and a 16 bit resolution. The speech signals were re-sampled to 16 kHz. All of the patients were

asked to read 97 words [1], which contain every phoneme of the German language in different positions within the words. The age of the CI users ranges from 33 up to 78 years old (57.2 ± 12.2). The age of the healthy speakers ranges from 31 up to 62 years old (44.2 ± 9.3).

2.2 Segmentation

The speech of the CI users is evaluated considering acoustic features extracted from CV/VC transitions formed with different phoneme groups of the standard German consonant system (Table 1). In order to obtain the time stamps of the phonemes in the

Table 1. Phoneme groups considered in this study.

Group	IPA Transcription
Nasals	/n/, /m/, /ŋ/
Sibilants	/s/, /ʃ/, /z/, /ʒ/
Fricatives	/f/, /v/, /j/, /ɕ/, /h/
Voiced stops	/b/, /d/, /g/
Voiceless stops	/p/, /t/, /k/

recordings, the BAS CLARIN web service is used [5]. This web service provides a forced alignment tool based on the Munich Automatic Segmentation system presented in [15]. The speech recordings are uploaded with their corresponding orthographic transcription to obtain the time stamps of the phonemes represented in SAMPA format. Then, the SAMPA segments are labeled according to Table 1. Additionally, short vowels, long vowels, and vowels that occur in unstressed position are labeled into one group. In the final step, the CV/VC transitions are extracted and grouped individually according to their phoneme label. Figure 2 summarizes this process.

2.3 Feature extraction

A Hamming window of 25 ms with a time step of 10 ms are applied from the beginning of the consonant (for CV transitions) or vowel (for VC transitions). The set of acoustic features includes 9 Perceptual Linear Predictive (PLP) coefficients and 13 Mel-Frequency Cepstral Coefficients (MFCCs). The mean, standard deviation, skewness, and kurtosis are computed from the descriptors, forming an 88-dimensional feature vector per speaker. Thus, there are 5 feature vectors per transition, one for each phoneme group.

2.4 Automatic classification

The automatic classification of postlingual CI users and HC is performed with a radial basis SVM with margin parameter C and a Gaussian kernel with parameter γ. C and

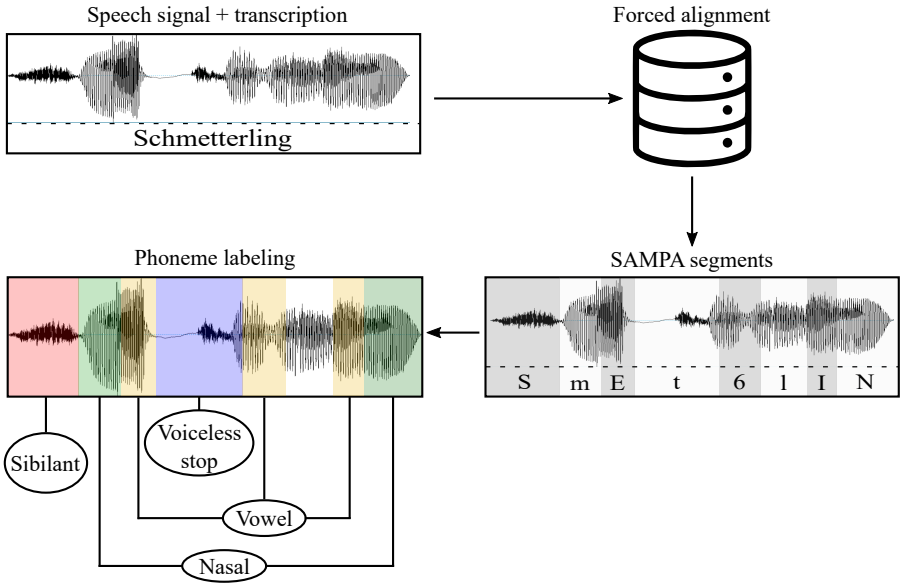

Fig. 2. Phoneme labeling procedure. In the figure, the German word "Schmetterling" contains two VC transitions (vowel-voiceless stop, vowel-nasal) and two CV transitions (voiceless stop-vowel, nasal-vowel)

γ are optimized through a grid search with $10^{-4} < C < 10^4$ and $10^{-6} < \gamma < 10^3$. The selection criterion is based on the performance obtained in the training stage. The SVM is tested following a 10-fold cross validation strategy. The performance of the system is evaluated by means of the accuracy (Acc), which measures the proportion of speakers that were assigned to the correct group by the system, the sensitivity (Sen), which measures the proportion of speakers correctly assigned to the CI users group, and the specificity (Spe), which measures the proportion of speakers correctly assigned to the HC group. Additionally, the Area Under the ROC Curve (AUC) is considered. The values of the AUC range from 0.0 up to 1.0, were 1.0 means a perfect system. The AUC is interpreted as follows: AUC < 0.70 indicates poor performance, $0.70 \leq$ AUC < 0.80 is fair, $0.80 \leq$ AUC < 0.90 is good, and $0.90 \leq$ AUC < 1 is excellent [16].

3 Experiments and results

Table 2 shows the obtained results for the classification of CI users and HC speakers when only features extracted from CV transitions are considered for training. It can be observed that the best performance is obtained with the sibilant-to-vowel transitions (Acc = 93%, AUC = 0.94), which confirms previous findings regarding the production of sibilant sounds in postlingual CI users [12]. Also, a difference can be observed when comparing the performance of voiceless-stop-to-vowel (Acc = 83%, AUC = 0.87) and voiced-stop-to-vowel (Acc = 70%, AUC = 0.73) transitions. In this case, the highest

Table 2. Results for the automatic classification between CI users vs HC speakers, considering CV transitions. Acc: Accuracy. Sen: Sensitivity. Spe: Specificity. AUC: Area under the ROC curve.

Transition	Phone group	Acc (%)	Sen (%)	Spe (%)	AUC
CV	Voiceless stops	83	80	85	0.87
	Voiced stops	70	65	75	0.73
	Sibilants	93	95	90	0.94
	Fricatives	85	90	80	0.87
	Nasals	88	85	90	0.90

results were obtained with the voiceless stops compared to the voiced stops. These results can be explained considering the study presented in [17]. The authors suggest that voiceless stop consonants require a more complex timing in coordinating the upper and laryngeal articulators than voiced stop consonants, which may be produced by simultaneous action of these articulators. Additionally, good results were also achieved for nasal-to-vowel (Acc = 88%, AUC = 0.90) and fricative-to-vowel (Acc = 85%, AUC = 0.87) transitions. Table 3 shows the obtained results for the classification of CI users and HC speakers when only features extracted from VC transitions are considered for training. In general, it can be observed that the classification results are lower than those presented in Table 2. For VC transitions, the highest results were obtained with fricative-to-vowel (Acc = 80%, AUC = 0.84) and sibilant-to-vowel (Acc = 78%, AUC = 0.85) transitions. Fair results were also obtained with voiced-stop-to-vowel transitions, however, from the sensitivity measure we can observe that the system is not able to identify postlingually deafened CI users properly (Sen = 60%).

Table 3. Results for the automatic classification between CI users vs HC speakers, considering VC transitions. Acc: Accuracy. Sen: Sensitivity. Spe: Specificity. AUC: Area under the ROC curve.

Transition	Phone group	Acc (%)	Sen (%)	Spe (%)	AUC
VC	Voiceless stops	68	70	65	0.74
	Voiced stops	75	60	90	0.71
	Sibilants	78	85	70	0.85
	Fricatives	80	80	80	0.84
	Nasals	63	65	60	0.63

4 Conclusions

In this paper we presented a study to investigate the use of acoustic features extracted from CV/VC transitions to detect speech problems in postlingually deafened CI users.

In order to do this, the transitions were grouped individually according to the manner of articulation of the consonants, i.e, voiceless stops, voiced stops, sibilants, fricatives, and nasals. According to the results, CV transitions prove to be more suitable than the VC transitions to detect changes in the speech of the patients in comparison to a group of HC speakers. Furthermore, the obtained results were similar to previous findings which are related to consonant production problems in CI users. The highest classification accuracy was obtained with features extracted from sibilant-to-vowel transitions (Acc = 93%), which indicates that there are differences between the production by CI users and HC controls. Additionally, good classification results were obtained with features from fricative-to-vowel (Acc = 85%), nasal-to-vowel (Acc = 88%), and voiceless-stop-to-vowel transitions (Acc = 93%). These results motivate us to implement this approach for the longitudinal monitoring of CI users. We are aware of a mismatch regarding the age and sex in CI and HC. Currently, we are collecting more HC. However, we don't expect the outcome of the experiments to change, i.e., that CV are better than VC and that sibilant-to-vowel transitions provide the best discrimination. The long term goal of this study is to provide to the expert clinicians with additional information that could be used to help the patients with their speech therapy. Future work will include more speech tasks such as text reading, rapid repetition of syllables, and sentence reading.

Acknowledgments

The authors acknowledge to the Training Network on Automatic Processing of PAthological Speech (TAPAS) funded by the Horizon 2020 programme of the European Commission. Tomás Arias-Vergara is under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS.

References

1. Fox-Boyer, A.: PLAKSS: Psycholinguistische Analyse kindlicher Sprechstörungen. Swets Test Services (2002)
2. Guenther, F.H., Perkell, J.S.: A neural model of speech production and its application to studies of the role of auditory feedback in speech. In: Maassen, B., Kent, R., Peters, H., van Lieshout, P., Hulstijn, W. (eds.) Speech Motor Control: In Normal and Disordered Speech, chap. 2, pp. 29–49. Oxford University Press, Great Clarendon Street, Oxford OX2 6DP (2004)
3. Hassan, S.M., Malki, K.H., Mesallam, T.A., Farahat, M., Bukhari, M., Murry, T.: The Effect of Cochlear Implantation on Nasalance of Speech in Postlingually Hearing-Impaired Adults. Journal of Voice 26(5), 669.e17 – 669.e22 (2012)
4. Hudgins, C.V., Numbers, F.C.: An investigation of the intelligibility of the speech of the deaf. Genetic psychology monographs (1942)
5. Kisler, T., Reichel, U., Schiel, F.: Multilingual processing of speech via web services. Computer Speech & Language 45, 326–347 (2017)
6. Lane, H., Webster, J.W.: Speech deterioration in postlingually deafened adults. The Journal of the Acoustical Society of America 89(2), 859–866 (1991)
7. Lane, H., Wozniak, J., Matthies, M., Svirsky, M., Perkell, J.: Phonemic resetting versus postural adjustments in the speech of cochlear implant users: An exploration of voice-onset time. The Journal of the Acoustical Society of America 98(6), 3096–3106 (1995)

8. Langereis, M., Dejonckere, P., Van Olphen, A., Smoorenburg, G.: Effect of cochlear implantation on nasality in post-lingually deafened adults. Folia phoniatrica et logopaedica 49(6), 308–314 (1997)
9. Leder, S.B., Spitzer, J.B.: A perceptual evaluation of the speech of adventitiously deaf adult males. Ear and hearing 11(3), 169–175 (1990)
10. Leder, S.B., Spitzer, J.B., Kirchner, J.C.: Speaking fundamental frequency of postlingually profoundly deaf adult men. Annals of Otology, Rhinology & Laryngology 96(3), 322–324 (1987)
11. Liberman, A., et al.: Some cues for the distinction between voiced and voiceless stops in initial position. Language and speech 1(3), 153–167 (1958)
12. Neumeyer, V., Schiel, F., Hoole, P.: Speech of cochlear implant patients: An acoustic analysis of sibilant production. In: ICPhS (2015)
13. Orozco-Arroyave, J.: Analysis of speech of people with Parkinson's disease. Logos-Verlag, Berlin, Germany (2016)
14. Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Perrier, P., Vick, J., Wilhelms-Tricarico, R., Zandipour, M.: A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. Journal of Phonetics 28(3), 233–272 (2000)
15. Schiel, F.: Automatic phonetic transcription of non-prompted speech. In: Proceedings of ICPhS. pp. 607–610 (1999)
16. Swets, J.A., et al.: Psychological science can improve diagnostic decisions. Psychological science in the public interest 1(1), 1–26 (2000)
17. Tobey, E.A., Pancamo, S., Staller, S.J., Brimacombe, J.A., Beiter, A.L.: Consonant production in children receiving a multichannel cochlear implant. Ear and Hearing 12(1), 23–31 (1991)
18. Ubrig, M.T., Goffi-Gomez, M.V.S., Weber, R., Menezes, M.H.M., Nemr, N.K., Tsuji, D.H., Tsuji, R.K.: Voice analysis of postlingually deaf adults pre-and postcochlear implantation. Journal of Voice 25(6), 692–699 (2011)