INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Phone-attribute posteriors to evaluate the speech of cochlear implant users

T. Arias-Vergara[1,2,3], *J.R. Orozco-Arroyave*[1,2], *M. Cernak*[4], *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
[4]Logitech, Laussane, Switzerland

tomas.ariasvergara@lme.de

Abstract

People with pre- and postlingual onset of deafness, i.e, age of occurrence of hearing loss, often present speech production problems even after hearing rehabilitation by cochlear implantation. In this paper, the speech of 20 prelinguals (aged between 18 to 71 years old), 20 postlinguals (aged between 33 to 78 years old) and 20 healthy control (aged between 31 to 62 years old) German native speakers are analyzed considering phone-attribute features extracted with pre-trained Deep Neural Networks. Speech signals are analyzed with reference to the manner of articulation of consonants according to 5 groups: nasals, sibilants, fricatives, voiced-stops, and voiceless-stops. According to the results, it is possible to detect alterations in the consonant production of CI users when compared with healthy speakers. A comprehensive evaluation of speech changes of CI users will help in the rehabilitation after deafening.

Index Terms: Hearing loss, Acoustic analysis, onset of deafness, Cochlear implant.

1. Introduction

The age of occurrence of deafness has an impact in speech production and understanding. On the one hand, when hearing loss occurs before speech acquisition (prelingual onset of deafness), a decreased speech intelligibility can be caused due to the fact that speakers have never monitored their own speech [1]. On the other hand, if hearing loss occurs after speech acquisition (postlingual onset of deafness), speech impairments are caused by the lack of sufficient and stable auditory feedback, but the person was able to correctly monitor his/her speech before deafening [2]. Furthermore, 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) [3, 4, 5]. 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 specific deficits of speech would be known the rehabilitation might be adequately addressed to each group. Until now, few studies have considered analysis of speech in pre- and postlingual CI users. In [6], a study was presented considering speech recordings of 10 CI users (5 prelingual) and 10 age-matched healthy controls. Articulation analysis was performed computing the vowel space of the German vowels /a/, /e/, /i/, /o/, and /u/. The authors reported a reduction of the vowel space area for the CI users respect to the healthy speakers. Recently in [7], significant differences were found in people with pre- and postlingual hearing loss. The authors performed a perceptual evaluation of 83 CI users (19 prelingual) in terms of manner, place, and type of articulation. The authors reported that the prelingual group made more articulation errors than the postlingual group and that the pattern was different. The most affected phonemes were sibilants (/s/, /z/, and /ʃ/) and stops (/p/, /b/, /t/, /d/, /k/, and /g/). Prelingually deafened people have difficulties learning how to speak intelligibly. On the other hand, the speech of postlingually deafened people is intelligible but it can sound abnormal when there is not sufficient auditory feedback [8]. In the literature only one study had considered automatic speech analysis of pre- and postlingual CI users. In [9], a study was presented to evaluate the speech intelligibility of CI users. Speech recordings from 50 healthy speakers and 50 CI users (14 prelingual) were considered. An Automatic Speech Recognition (ASR) system was used to compute the Word Recognition (WR) rate. The authors reported higher WR values for the postlingual group compared to the prelingual group. The alteration of articulatory movements in hearing impaired people can be explained with the neural model of speech production (DIVA model) proposed in [10]. In the model, articulatory movements are planned to achieved auditory goals. During planning, the motor system uses 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. As described in [11], after cochlear implantation, the user may notice differences between the sounds perceived and the sounds produced. If this is the case, then the patient will move the articulators in order to produce a speech sound similar to the sound perceived. For instance, previous work suggests that sibilant production differs between CI users and healthy speakers 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]. In terms of speech production it is clear that there are differences between pre- and postlingual deafened CI users. Thus, it is expected to detect these differences using automatic

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2144

classification. However, is not the aim of this study to differentiate between CI users and healthy speakers. Instead, we propose a methodology to detect speech problems in CI users automatically. Since several articulatory settings are require to produce different speech sounds, the acoustic analysis of consonant groups can be associated to specific motor control problems. This paper investigates the use of phone-attribute features to detect speech problems in pre- and postlingual deafened CI users. In order to do this, phonemes are detected from the recordings and grouped into nasals, sibilants, fricatives, voiced-stops, or voiceless-stops. Acoustic features and phone-attribute posteriors are computed from the speech signals and the consonants are extracted to perform automatic classification of CI users and healthy speakers. In the long run we want to develop supporting therapy technology that can integrate speech perception and production analysis in order to perform an adapted speech therapy.

2. Materials and methods

Figure 1 shows the methodology implemented in this work. First, forced alignment is performed over the speech recordings uttered by each speaker. Next, the phonemes are labeled according to five consonant groups: nasals, sibilants, fricatives, voiced-stops, and voiceless-stops. Then, acoustic features and phone-attribute posteriors are extracted from the recordings and the consonants are grouped according to the phonemes groups listed before. The set of features computed includes the duration of each consonant, Perceptual Linear Predictive (PLP) coefficients, Mel-Frequency Cepstral Coefficients (MFCCs), and phone-attribute posteriors, which are computed using a Deep Neural Network (DNN) approach. After feature extraction, a three-class Support Vector Machine (SVM) is considered for the automatic classification between CI users and healthy controls (HC). Each stage of the methodology is described in more detail in the following sections.

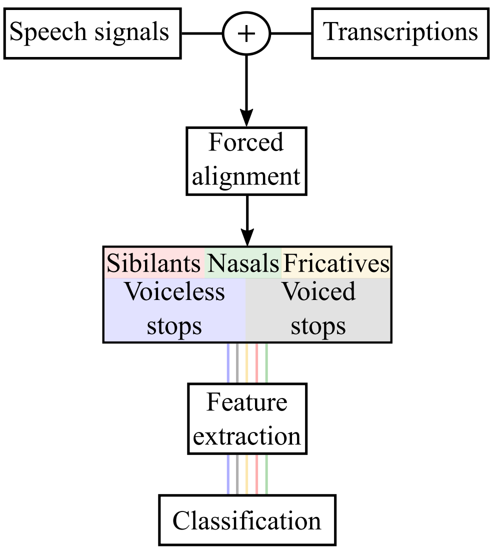

Figure 1: *Methodology implemented in this study.*

2.1. Data

Standardized speech recordings of 20 prelingual (PRE), 20 postlingual (POST) deafened CI users, and 20 HC German native speakers were considered for the tests. Detailed information of the speakers is presented in Table 1. The speech signals were captured in noise-controlled conditions at the Clinic of the Ludwig-Maximillians 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 [13], which contain every phoneme of the German language in different positions within the words. Figure 2 shows the age distribution of the speakers considered for the experiments. Since there is a age mismatch in the speaker groups, a regression approach is considered to validate the classification results (Section 2.5).

Table 1: *Information of the speakers. PRE: prelingual CI users. POST: postlingual CI users. HC: healthy controls. μ: mean. σ: standard deviation. m: Number of male speakers. f: Number of female speakers.*

	PRE	**POST**	**HC**
	m: 6 f: 14	m: 4 f: 16	m: 11 f: 9
Range of age [years]	18 - 71	33 - 78	31 - 62
Age [years] ($\mu \pm \sigma$)	35.6 ± 18.5	57.2 ± 12.2	44.2 ± 9.3

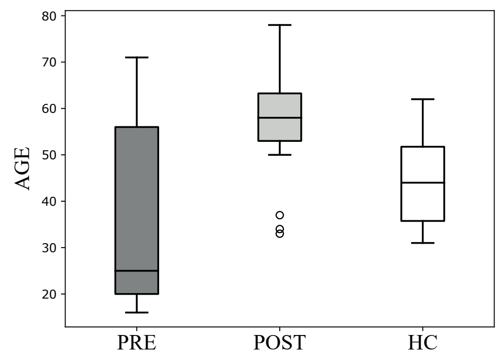

Figure 2: *Age distribution of the speakers considered in this study. PRE: prelingual deafened CI users, POST: postlingual deafened CI users, HC: healthy controls.*

2.2. Segmentation

Every phoneme in the speech recordings is detected automatically using the BAS CLARIN web service, which allows to perform forced alignment [14]. The speech recordings are uploaded with their corresponding orthographic transcription to obtain the time stamps of the phonemes represented in SAMPA format. Then, the consonants are assigned to five phoneme groups which are formed considering the German consonant system. Table 2 shows the phonemes and groups used in this study.

Table 2: *Consonant groups considered in this study.*

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

2.3. Feature extraction

Each recording is divided into frames of 25 ms length, with a time-step of 10 ms. Hamming windowing is applied to every frame before feature extraction. The sequence of frames is converted into a sequence of feature vectors $X = \{x_1, ..., x_n, ..., x_N\}$, where N is the number of frames extracted from the speech signal. Then, the time-stamps of the phonemes are considered to extract the feature vectors associated with the consonants listed in Table 2. This procedure is repeated for the 97 words uttered by each speaker. Figure 3 summarizes the procedure. The phone-attribute posteriors are extracted using the deep learning approach presented in [15]. In that work, the posteriors are computed by a bank of parallel DNN, which converts a sequence of acoustic features $Z = \{z_1, ..., z_n, ..., z_N\}$ into a sequence of vectors $P = \{p_1, ..., p_n, ..., p_N\}$. The vector $P = \{p_1, ..., p_n, ..., p_N\}$ consists of K posterior probabilities. Each DNN estimates the posterior p_n^k as the probability of occurrence of the k-th phone-attribute feature. In this work, a set of 4 pre-trained DNNs are used to extract the posteriors. The DNNs were trained with the librispeech corpus ([16]) following the sound pattern of English. The phone-attribute features included are: consonantal, which indicates sounds where there is an obstruction of the vocal tract. Continuant, which differentiates non-plosive from plosive sounds. Strident, which refers to sounds with more energy in higher frequency bands. Nasal, which a lowered velum, where the air escape through the nose. The following procedure is performed for every consonant group (Table 2): Each phone-attribute posterior is used as a weight to multiply 9 PLPs and 13 MFCCs, forming a 22 dimensional feature vector per phoneme posterior. Then, the feature vectors are concatenated together with the duration of the phoneme to form a 89-dimensional feature vector. Four functionals (mean, standard deviation, kurtosis, and skewness) are computed to form a 356-dimensional feature vector per speaker.

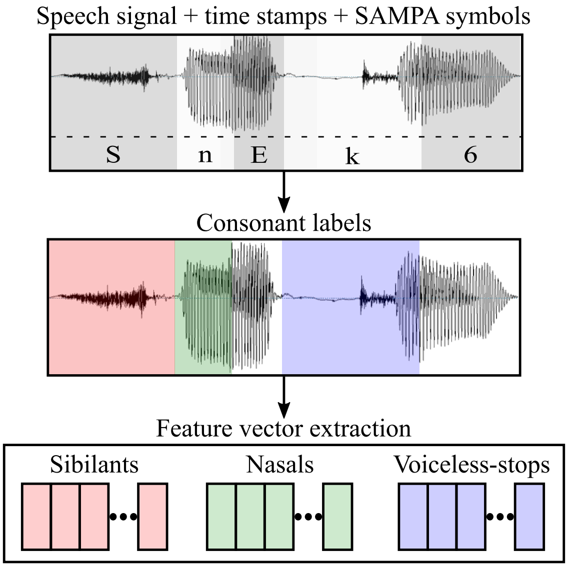

Figure 3: *Feature extraction procedure. In the figure, the German word "Schnecke" contains one sibilant (/ʃ/), one nasal (/n/), and one voiceless-stop (/k/).*

2.4. Automatic classification

A radial basis SVM with margin parameter C and a Gaussian kernel with parameter γ is used for automatic classification. C and γ are optimized through a grid-search with $10^{-4} \leq C \leq 10^4$ and $10^{-6} \leq \gamma \leq 10^3$. 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 three-class SVM is evaluated by means of the precision, recall, and F1-score.

2.5. Regression analysis

As described in Section 2.1, there is an age mismatch in CI users and HC speakers. Thus, a linear Support Vector Regressor with an ε-insensitive loss function (ε–SVR) is trained in order to determine whether the classification results are biased by the age of the speakers. The parameters of the ε–SVR C, γ and ε are optimized in a grid search with $10^{-4} \leq C \leq 10^3$ and $10^{-4} \leq \varepsilon \leq 10^3$. The selection criterion is based on the performance obtained in the training stage. The ε–SVR is tested following a 5-fold cross validation strategy. The performance is evaluated using the Pearson's correlation coefficient (ρ) and the Mean Absolute Error (MAE) between the predicted values and the age of the speakers. Pearson's ρ varies between -1 and 1 and is interpreted as follows: $0.00 \leq |\rho| < 0.20$ indicates "very weak" correlation, $0.20 \leq |\rho| < 0.40$ is "weak", $0.40 \leq |\rho| < 0.60$ is "moderate", $0.60 \leq |\rho| < 0.80$ is "strong", and $0.80 \leq |\rho| \geq 1$ is "very strong" [17].

3. Experiments and results

Table 3 shows the obtained results for the automatic classification of PRE, POST, and HC speakers. In general, we can observe that there are differences comparing the speech of CI users(PRE and POST) with HC speakers. These differences were found in sibilants (HC:F1-score=82%), fricatives (HC:F1-score=76%), voiced-stops (HC:F1-score=74%), voiceless-stops (HC:F1-score=74%). In sibilant sounds, an alteration may be associated with a lower spectral resolution for consonant production in higher frequencies [12]. Alterations in stop sounds production has been also reported in previous work ([18, 19]), however, these studies evaluated voicing contrast considering the voice onset time. Furthermore, we can observe that the performance is better for voiceless-stops compared to voiced-stops. In [20], the authors suggest that voiceless-stop consonants require a more complex timing in coordinating the upper and laryngeal articulators than voiced-stop consonants. This timing may be produced by simultaneous action of the upper and laryngeal articulators. No differences in speech production between patients and controls were found using features extracted from the nasal group. However, nasal consonant production problems can occur due to a lack of coordination in articulatory movements leading to a nasalization in speech [21, 22, 23], which can be detected with features extracted from nasal-to-vowel or nasal-to-consonant transitions. Regarding the onset of deafness, no clear differences were found with the proposed approach. The best results were obtained with the sibilants. Speech alterations are detected in the postlingual group (F1-score=70%), but not in the prelinguals (F1-score=47%). In order to verify whether the classification results are biased by the age mismatch, phone-attribute features are used to train a linear ε–SVR. Table 4 shows the obtained results. Note that only the voiced-stop group shows a "moderate", which means that there is a significant correlation between

Table 3: *Classification results with phone-attribute posteriors, PLPs, and MFCCs. PRE: prelingual deafened CI users. POST: postlingual deafened CI users. HC: healthy controls. Avg: Average. Results in %.*

Consonant group	Class	Precision	Recall	F1-score
	PRE	61	55	58
Nasals	POST	48	50	49
	HC	48	50	49
	Avg	52	52	52
	PRE	70	35	47
Sibilants	POST	62	80	70
	HC	75	90	82
	Avg	69	68	66
	PRE	40	20	27
Fricatives	POST	50	50	50
	HC	63	95	76
	Avg	51	55	51
	PRE	40	40	40
Voiced-stops	POST	35	30	32
	HC	70	80	74
	Avg	48	50	49
	PRE	57	65	60
Voiceless-stops	POST	64	45	53
	HC	70	80	74
	Avg	63	63	63

the phone-attribute features extracted from the voiced-stop consonants and the age of the speakers. Furthermore, "weak" correlations were found for the voiceless-stop and nasal consonant groups, and "very weak" correlations were found in the sibilant and fricative groups. Further experiments were performed without the phone-attribute posteriors, i.e, only the phoneme duration, 9 PLPs and 13 MFCCs were considered for feature extraction. Table 5 shows the obtained results. We can observe that for sibilants, the F1-score for the postlinguals is lower (F1-score=58%) compared with the results obtained when phone-attribute posteriors are considered (F1-score=70%). Similar results were found in the voiceless-stops. In this case, the performance improves from F1-score=43% to F1-score=60% in the prelingual group, when the phone-attribute posteriors are considered during feature extraction.

Table 4: *Pearson ρ and MAE between the age of the speakers and the age estimated with the linear ε–SVR.*

Consonant group	Pearson ρ	MAE
Nasals	0.30	14.2
Sibilants	0.26	14.7
Fricatives	0.23	14.1
Voiced-stops	0.43	13.9
Voiceless-stops	0.35	14.0

4. Conclusions

Speech of CI users shows differences in comparison with normal hearing persons and in between those with prelingual and postlingual hearing loss. In this paper we presented a study to investigate the use of phone-attribute features to detect speech production referring to the most frequent differences in CI users. Phone-attribute posteriors were computed considering a deep learning approach. Although, the DNNs used to obtain the posteriors are pre-trained with English speakers, phone-attribute posteriors proved to be useful to detect speech problems. This was observed when acoustic features with and without phone-attribute posteriors were used for automatic classification. In order to detect speech production problems, the phonemes 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, it is possible to detect speech production problems in CI users. Particularly, when sibilant consonants are considered. Differences between healthy speakers and CI users were also found in voiceless-stops and voiced-stops. However, this approach could be extended especially by using other methods focusing on voice onset time, which have been used in previous studies to detect voicing problems in hearing impaired people. In the presented procedure, no differences in speech were found in the nasal sounds. Therefore, it could be necessary to consider other approaches such as acoustic analysis of nasal-to-vowel transitions. Currently, we are testing recurrent neural networks trained on German databases to compute the phone-attribute posteriors using the consonant groups considered in this study. Further, the data collection is still ongoing in order to include more age-matched HC controls and patients.

Table 5: *Classification results with only PLPs and MFCCs. PRE: prelingual deafened CI users. POST: postlingual deafened CI users. HC: healthy controls. Avg: Average. Results in %*

Consonant group	Class	Precision	Recall	F1-score
	PRE	37	35	36
Nasal	POST	31	25	28
	HC	40	50	44
	Avg	36	37	36
	PRE	55	60	57
Sibilants	POST	61	55	58
	HC	80	80	80
	Avg	65	65	65
	PRE	45	25	32
Fricatives	POST	44	60	51
	HC	82	90	86
	Avg	57	58	56
	PRE	53	45	49
Voiced-stops	POST	56	50	53
	HC	64	80	71
	Avg	57	58	57
	PRE	47	40	43
Voiceless-stops	POST	52	65	58
	HC	78	70	74
	Avg	59	58	58

5. 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. The authors also thanks to CODI from University of Antioquia (grant Numbers PRV16-2-01 and 2015–7683).

6. References

[1] C. R. Smith, "Residual hearing and speech production in deaf children," *Journal of Speech, Language, and Hearing Research*, vol. 18, no. 4, pp. 795–811, 1975.

[2] S. B. Leder and J. B. Spitzer, "A perceptual evaluation of the speech of adventitiously deaf adult males." *Ear and hearing*, vol. 11, no. 3, pp. 169–175, 1990.

[3] C. V. Hudgins and F. C. Numbers, "An investigation of the intelligibility of the speech of the deaf," *Genetic psychology monographs*, 1942.

[4] M. Langereis, P. Dejonckere, A. Van Olphen, and G. Smoorenburg, "Effect of cochlear implantation on nasality in post-lingually deafened adults," *Folia phoniatrica et logopaedica*, vol. 49, no. 6, pp. 308–314, 1997.

[5] S. B. Leder, J. B. Spitzer, and J. C. Kirchner, "Speaking fundamental frequency of postlingually profoundly deaf adult men," *Annals of Otology, Rhinology & Laryngology*, vol. 96, no. 3, pp. 322–324, 1987.

[6] V. Neumeyer *et al.*, "An acoustic analysis of the vowel space in young and old cochlear-implant speakers," *Clinical linguistics & phonetics*, vol. 24, no. 9, pp. 734–741, 2010.

[7] D. Splitthoff *et al.*, "Consonant articulation errors of adult and adolescent cochlear implant recipients with respect to onset and duration of deafness," *International Journal of Speech-Language Pathology (Under review)*, 2018.

[8] J. Perkell *et al.*, "The sensorimotor control of speech production," in *Proceedings of the First International Symposium on Measurement, Analysis and Modeling of Human Functions*, 2001, pp. 21–23.

[9] S. Ruff, T. Bocklet, E. Nöth, J. Müller, E. Hoster, and M. Schuster, "Speech Production Quality of Cochlear Implant Users with Respect to Duration and Onset of Hearing Loss," *ORL*, vol. 79, no. 5, pp. 282–294, 2017.

[10] F. H. Guenther, J. S. Perkell, B. Maassen, R. Kent, H. Peters, P. van Lieshout, and W. Hulstijn, "A neural model of speech production and its application to studies of the role of auditory feedback in speech," *Speech motor control in normal and disordered speech*, pp. 29–49, 2004.

[11] H. Lane, J. Wozniak, M. Matthies, M. Svirsky, and J. Perkell, "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*, vol. 98, no. 6, pp. 3096–3106, 1995.

[12] V. Neumeyer, F. Schiel, and P. Hoole, "Speech of cochlear implant patients: An acoustic analysis of sibilant production." in *ICPhS*, 2015.

[13] A. Fox-Boyer, *PLAKSS: Psycholinguistische Analyse kindlicher Sprechstörungen*. Swets Test Services, 2002.

[14] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[15] M. Cernak *et al.*, "Phonological vocoding using artificial neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4844–4848.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[17] R. Taylor, "Interpretation of the correlation coefficient: a basic review," *Journal of diagnostic medical sonography*, vol. 6, no. 1, pp. 35–39, 1990.

[18] H. Lane and J. W. Webster, "Speech deterioration in postlingually deafened adults," *The Journal of the Acoustical Society of America*, vol. 89, no. 2, pp. 859–866, 1991.

[19] L. Scarbel, A. Vilain, H. Loevenbruck, and S. Schmerber, "An acoustic study of speech production by French children wearing cochlear implants," in *3rd Early Language Acquisition Conference*, 2012.

[20] E. A. Tobey, S. Pancamo, S. J. Staller, J. A. Brimacombe, and A. L. Beiter, "Consonant production in children receiving a multichannel cochlear implant." *Ear and Hearing*, vol. 12, no. 1, pp. 23–31, 1991.

[21] S. M. Hassan, K. H. Malki, T. A. Mesallam, M. Farahat, M. Bukhari, and T. Murry, "The Effect of Cochlear Implantation on Nasalance of Speech in Postlingually Hearing-Impaired Adults," *Journal of Voice*, vol. 26, no. 5, pp. 669.e17 – 669.e22, 2012.

[22] R. M. Uchanski and A. E. Geers, "Acoustic characteristics of the speech of young cochlear implant users: A comparison with normal-hearing age-mates," *Ear and hearing*, vol. 24, no. 1, pp. 90S–105S, 2003.

[23] R. B. Monsen, "Toward measuring how well hearing-impaired children speak," *Journal of Speech and Hearing Research*, vol. 21, no. 2, pp. 197–219, 1978.