# Multi-channel Convolutional Neural Networks For Automatic Detection of Speech Deficits in Cochlear Implant Users.

T. Arias-Vergara[1,2,3], J.C. Vasquez-Correa[1,2], S. Gollwitzer[3], J.R. Orozco-Arroyave[1,2], 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.** This paper proposes a methodology for automatic detection of speech disorders in Cochlear Implant users by implementing a multi-channel Convolutional Neural Network. The model is fed with a 2-channel input which consists of two spectrograms computed from the speech signals using Mel-scaled and Gammatone filter banks. Speech recordings of 107 cochlear implant users (aged between 18 and 89 years old) and 94 healthy controls (aged between 20 and 64 years old) are considered for the tests. According to the results, using 2-channel spectrograms improves the performance of the classifier for automatic detection of speech impairments in Cochlear Implant users.

**Keywords:** Speech processing · time-frequency analysis · Multi-channel CNN · Deep Learning · Cochlear Implants

## 1 Introduction

Speech disorders affect the communication ability of people affected by certain medical conditions such as hearing loss, laryngeal and oral cancer, neurodegenerative diseases such as Parkinson's disease, and others. For the case of hearing loss, there are different treatments available depending on the degree and type of deafness. Cochlear Implants (CIs) are the most suitable devices when hearing aids no longer provide sufficient auditory feedback. However, CI users often experience alteration in speech even after rehabilitation, such as decreased intelligibility and changes in terms of articulation [1]. Thus, the development of computer aided systems will contribute to support the diagnosis and monitoring of speech. In the literature, few studies have addressed acoustic analysis of speech of CI users by implementing machine learning methods. In [2] speech intelligibility of 50 CI users is evaluated using an automatic speech recognition system and compared with 50 Healthy Controls (HC). Recently in [3] automatic

classification using Support Vector Machines (SVM) between 20 CI users and 20 healthy speakers was performed in order to evaluate articulation disorders considering acoustic features. For the case of pathological speech detection, CNNs have outperformed classical machine learning methods [4–6]. In these studies, the conventional method is to perform time-frequency analysis by computing spectrograms over the speech signals to feed the CNNs with single channel inputs. However, using one channel as input may limit the potential of the model to learn more complex representations of speech signals.

In this study we propose a deep learning-based approach for the automatic detection of disordered speech in postlingually deafened CI users, i.e, when hearing loss occurs after speech acquisition. The method consists of 2-channel spectrograms as input to a CNN. Time-frequency analysis is performed considering Mel-scaled and Gammatone spectrograms, which are computed from short-time segments extracted from the recordings. These segments are defined as the transitions from voiceless to voiced sounds (onset) and voiced to voiceless sounds (offset). Our main hypothesis is that using the spectrograms as a 2-channel input will allow the CNN to complement the information from the two time-frequency representations. On the one hand, Mel-based features have been established as the standard feature set for different speech and audio processing applications. On the other hand, previous studies have shown that Gammatone-based features are more robust to noise compared with Mel features [7].

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

### 2.1    Data

Standardized speech recordings of 107 CI users (56 male, 51 female) and 94 HC (46 male, 48 female) German native speakers are considered for the experiments. All of the CI users and 31 of the 94 healthy speakers were recorded at the Clinic of the Ludwig-Maximilians University in Munich (LMU), with a sampling frequency of 44.1 kHz and a 16 bit resolution. The recordings of the remaining 63 HC speakers were extracted from the PhonDat 1 (PD1) corpus from the Bavarian Archive For Speech Signals (BAS), which is freely available for European academic users[1]. For this corpus, the subjects were labeled as "old" and "young", however, the age of the speakers is not included in the description of the dataset. Speech recordings from the BAS corpus have a sampling frequency of 16 kHz. The mismatch in the acoustic conditions of the BAS corpus and our recordings is addressed in Section 2.2. The speech recordings include the reading of *Der Nordwind und die Sonne* (*The North Wind and the Sun*) text. Information about the subjects considered in this study is presented in Table 1.

---

[1] http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B

**Table 1.** Information of the speakers. HC-Clinic: Healthy speakers recorded in the clinic. HC-BAS: Healthy speakers extracted from the BAS repository. $\mu$: mean. $\sigma$: standard deviation

|  | CI | | HC-Clinic | | HC-BAS | |
|---|---|---|---|---|---|---|
|  | Male | Female | Male | Female | Male | Female |
| N. Speakers | 56 | 51 | 11 | 20 | 35 | 28 |
| Range of age | 18-89 | 28-84 | 26-53 | 20-64 | - | - |
| Age ($\mu \pm \sigma$) | 65±16 | 62±15 | 35±9 | 37±13 | - | - |

## 2.2   Preprocessing

The first step is to remove any possible DC offset induced by the microphone and to re-scale the amplitude of the speech signals between -1 and 1. Then, a noise reduction method and a compression technique are applied to normalize the acoustic conditions of the recordings from the clinic and BAS. Then, onset and offset transitions are extracted to model speech disorders in CI users. The details of the methods implemented are as follows:

**Noise reduction -** Background noise is removed using the SoX codec[2]. The noise reduction algorithm is based on spectral gating, which consists in obtaining a profile of the background noise to enhance the quality of the audio. In order to get the profile, the Short-Time Fourier Transform (STFT) is computed over short-time frames extracted from a noisy signal (silence region from the recording to be denoise). Then, the mean power is computed over each point of the STFT in order to get thresholds per each frequency band. The STFT of the complete signal is calculated and the sounds with energies lower than the thresholds are attenuated for noise reduction. For more details regarding the implementation, please refer to the official SoX website.

**Compression -** The GSM full-rate compression technique is considered to normalize the acoustic conditions of the recordings from the clinic and the BAS repository [8]. The denoised speech signals are down-sampled to 8 kHz and the resolution is lowered to 13 bits, with a compression factor of 8. Additionally, a bandpass filter between 200 Hz and 3.4 kHz is applied in order to meet the specifications of a GSM transmission network. Figure 1 shows the STFT spectrograms of a speech recording before and after applying noise reduction and compression. The figures correspond to a speech segment of 600 ms extracted from the speech signal of one of the healthy speakers recorded in the clinic.

**Segmentation -** Speech signals are analyzed based on the automatic detection of onset and offset transitions, which are considered to model the difficulties of

---

[2] http://sox.sourceforge.net/
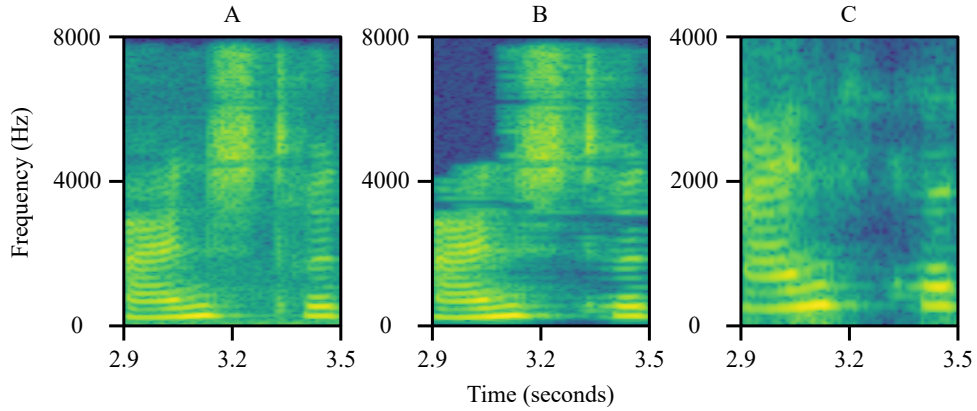
**Fig. 1.** Time-frequency representation of a segment from a speech signal. The figure shows (A) the original signal, (B) the signal after noise reduction, and (C) the signal after compression.

the patients to start/stop the movement of the vocal folds. The method used to identify the transitions is based on the presence of the fundamental frequency of speech (pitch) in short-time frames as it was shown in [9]. The transition is detected, and 80 ms of the signal are taken to the left and to the right of each border, forming segments with 160 ms length (Figure 2).
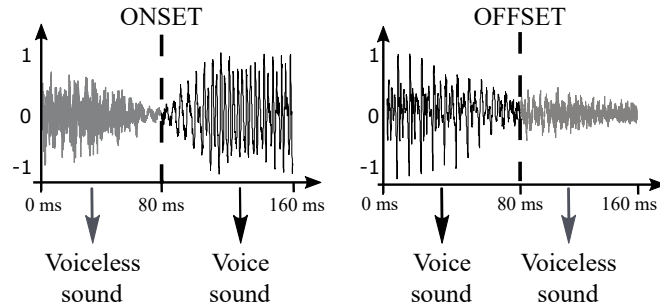


**Fig. 2.** ONSET and OFFSET transition frames.

### 2.3   Acoustic analysis

Acoustic features are extracted from the onset/offset transitions based on two different auditory filter banks. First, acoustic features are extracted by applying triangular filters on the Mel scale. Frequencies in Hz can be converted to Mel scale as:

$$M(f_{\text{Hz}}) = 1125 \ln(1 + f_{\text{Hz}}/700) \tag{1}$$

In the second approach, features are extracted using Gammatone filter banks, which are based on the cochlear model proposed in [10]. The model consists of an array of bandpass filters organized from high frequency at the base of the cochlea, to low frequencies at the apex (innermost part of the cochlea). The Gammatone filter bank is defined in the time domain by Equation 2 as:

$$g(t) = a t^{n-1} \exp(-2\pi b t) \cos(2\pi f_c t + \phi) \tag{2}$$

Where $f_c$ is the filter's center frequency in Hz, $\phi$ is the phase of the carrier in radians, $a$ is the amplitude, $n$ is the order of the filter, $b$ is the bandwidth in Hz, and $t$ is the time. The number of filters used for both Mel-scale and Gammatone based features is $n = 64$. The Gammatone filters are implemented following the procedure described in [11]. Features are extracted from the transitions using Hanning windows of 20 ms length with a time step of 5 ms.

### 2.4   Baseline model

Mel-Frequency Cepstral Coefficients (MFCCs) and Gammatone-Frequency Cepstral Coefficients (GFCC) are extracted by dividing the transitions into short-time segments $X = \{x_1, ..., x_n\}$. Then, the Mel/Gammatone filter bank is applied and the discrete cosine transform is calculated upon the logarithm of the energy bands using Equation 3.

$$\text{coef}[k] = 2 \sum_{i=0}^{n-1} x_f[i] \cos(\pi k (2i+1)/2n) \tag{3}$$

Where $k$ are the coefficients and $x_f$ is the resulting signal after applying the filter banks. In this work, 13 MFCCs (including the energy of the signal) and 12 GFCCs are considered. The mean, standard deviation, kurtosis, and skewness are computed from the descriptors. The automatic classification between CI users and HC speakers is performed with a radial basis SVM with margin parameter $C$ and a Gaussian kernel with parameter $\gamma$. $C$ and $\gamma$ are optimized through a grid-search up to powers of ten 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 implemented with scikit-learn [12].

### 2.5   Proposed model

Mel-scaled and the Gammatone spectrograms (Cochleagram) are computed from the onset/offset transitions by applying the filter banks described before. Then, the spectrograms are combined into a 2-channel tensor to fed the CNN, which is implemented using PyTorch [13]. From the documentation, it can be observed that the output of the convolutional layer for an input signal $(Bs, C_{\text{in}}, H, W)$ is described as:

$$h(Bs_i, C_{\text{out}_j}) = \text{bias}(C_{\text{out}_j}) + \sum_{k=0}^{C_{\text{in}}-1} \omega(C_{\text{out}}, k) * \text{input}(Bs_i, k) \tag{4}$$

Where $Bs$ is the batch size ($Bs = 100$), $\omega$ are the weights of the network, $C$ is the number of channels ($C = 2$) of the input tensor, $H$ is the height of the input signal ($H = 64$, number of filter banks), and $W$ is the width of the input signal ($W = 28$, number of frames in the onset/offset transitions). The architecture of the CNN implemented in this study is summarized in Figure 3. It consists of two convolutional layers, two max-pooling layers, dropout to regularize the weights, and two fully connected hidden layer followed by the output layer to make the final decision using a softmax activation function. The CNN is trained using the Adam optimization algorithm [14] with a learning rate of $\eta = 10^{-4}$. The cross–entropy between the training labels $y$ and the model predictions $\widehat{y}$ is used as the loss function. The size of the kernel in the convolutional layers is $k_c = 3x3$. For the pooling layers the kernel's size is $k_p = 2x2$. None of the network hyper-parameters are optimized in order to have comparable models across experiments.
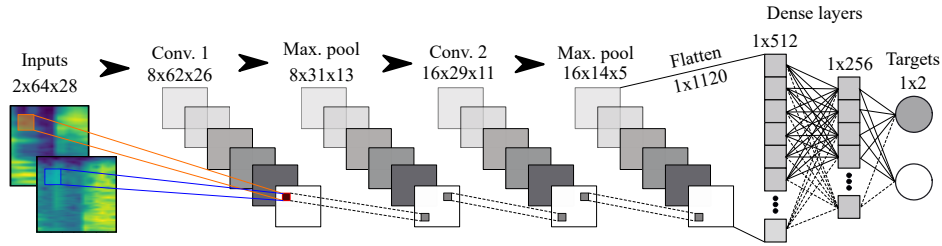


**Fig. 3.** Architecture of the CNN implemented in this study. The size of the kernel in the convolutional (Conv. $i$) and pooling layers (Max. pool) is 3x3 and 2x2, respectively.

## 3    Experiments and results

The SVMs and CNNs are tested following a 10-Fold Cross-validation strategy. The performance of the system is evaluated by means of the accuracy (Acc), sensitivity (Sen), and the specificity (Spe). The SVM and the multi-channel CNN are trained with features/spectrograms extracted from onset and offset transitions, individually. Table 2 shows the results obtained for the baseline model. It can be observed that the accuracies are higher for the offset transitions, when MFCCs and GFCCs are considered to train the SVMs individually, however, the best performance is achieved when the two feature sets are combined (Onset-Acc = 82.4%; Offset-Acc = 83.4%). Additionally, note that the sensitivity values are higher in all of the experiments. This can be explained considering that the speech of some of the CI users may be not affected, thus, it is closer to the speech of healthy speakers. Table 3 shows the results obtained with the proposed approach. In general, the accuracies of the CNNs are higher than those of the baseline model. This is mainly because the CNNs are able to classify more

**Table 2.** Classification results for the SVM trained with MFCCs and GFCCs features. **Acc**: Accuracy. **Sen**: Sensitivity. **Spe**: Specificity. Fusion: Combination of MFCCs and GFCCs features.

| Segment | Features | Acc (%) | Sen (%) | Spe (%) |
|---------|----------|---------|---------|---------|
|         | MFCC     | 80.6    | 85.2    | 76.0    |
| Onset   | GFCC     | 75.3    | 78.7    | 72.0    |
|         | Fusion   | 82.4    | 86.0    | 78.7    |
|         | MFCC     | 83.4    | 93.5    | 73.5    |
| Offset  | GFCC     | 82.8    | 88.8    | 76.9    |
|         | Fusion   | 83.4    | 91.7    | 75.2    |

CI users, which is not the case for the HC. As explained before not every CI users may present speech deficits, thus, it is not expected that the classifiers discriminate all of the speakers correctly. Nevertheless, it can be observed that the combination of the Mel-spectrogram and Cochleagram into a 2-channel tensor is suitable for the automatic detection of speech deficits. Additionally, note that this methodology is not restricted only to analyze speech of CI users, but it can be adapted to study other pathologies or to recognize other paralinguistic aspects from speech signals such as emotions.

**Table 3.** Classification results for the CNNs trained with Mel-spectrograms and Cochleagrams. **Acc**: Accuracy. **Sen**: Sensitivity. **Spe**: Specificity. **Fusion**: 2-channel spectrograms (Mel-spectrogram–Cochleagram).

| Segment | Inputs | Acc (%) | Sen (%) | Spe (%) |
|---------|--------|---------|---------|---------|
|         | Mel-spectrogram | 83.5 | 98.1 | 68.8 |
| Onset   | Cochleagram     | 85.4 | 95.4 | 75.4 |
|         | Fusion          | 86.8 | 98.2 | 75.4 |
|         | Mel-spectrogram | 85.9 | 96.4 | 75.5 |
| Offset  | Cochleagram     | 86.8 | 96.3 | 77.3 |
|         | Fusion          | 86.8 | 97.3 | 76.3 |

## 4   Conclusions

In this paper we presented a methodology for automatic detection of speech deficits in CI users using multi-channel CNNs. The method consists in combining two types of time-frequency representations into a 2-channel tensor which is used to fed a CNN. In order to do this, Mel-spectrograms and Cochleagrams are computed from onset and offset transitions extracted from the recordings of CI users and healthy speakers. Cepstral coefficients and SVM classifiers were

considered for comparison. According to the results, it is possible to differentiate between CI users and HC with accuracies of up to 86.8% when the multi-channel CNN is considered. We are aware of a mismatch regarding the age of the CI users and HC. Currently, we are collecting more HC, however, we don't expect the outcome of the experiments to change. Additionally, note that the multi-channel CNN may be suitable for other speech processing tasks such as emotion detection or as the feature stage for speech recognition. Future work should include more time-frequency analysis such as Perceptual Linear Prediction and the analysis of other pathologies in other to validate the proposed approach.

## Acknowledgments

## References

1. C. V. Hudgins and F. C. Numbers. An investigation of the intelligibility of the speech of the deaf. *Genetic psychology monographs*, 1942.
2. 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*, 79(5):282–294, 2017.
3. T. Arias-Vergara, J. R. Orozco-Arroyave, S. Gollwitzer, M. Schuster, and Elmar Nöth. Consonant-to-Vowel/Vowel-to-Consonant Transitions to Analyze the Speech of Cochlear Implant Users. *Accepted to be presented at International Conference on Text, Speech, and Dialogue*, 2019.
4. T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia. Dysarthric speech recognition using a convolutive bottleneck network. In *2014 12th International Conference on Signal Processing (ICSP)*, pages 505–509. IEEE, 2014.
5. Y. Takashima et al. Audio-Visual Speech Recognition Using Bimodal-Trained Bottleneck Features for a Person with Severe Hearing Loss. In *Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association*, pages 277–281, 2016.
6. J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth. Convolutional neural network to model articulation impairments in patients with parkinson's disease. In *Proceedings of the Eighteenth Annual Conference of the International Speech Communication Association*, pages 314–318, 2017.
7. X. Zhao and D. Wang. Analyzing noise robustness of mfcc and gfcc features in speaker identification. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 7204–7208. IEEE, 2013.
8. J. M. Huerta and R. M. Stern. Speech recognition from GSM codec parameters. In *Fifth International Conference on Spoken Language Processing*, 1998.

9. J. R. Orozco-Arroyave. *Analysis of Speech of People with Parkinson's Disease.* Logos Verlag Berlin, Germany, 2016.

10. R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In *Auditory physiology and perception*, pages 429–446. Elsevier, 1992.

11. M. Slaney et al. An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, 35(8), 1993.

12. F. Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

13. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. 2017.

14. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International conference on learning representation (ICLR)*, 2015.