ORIGINAL ARTICLE

Multi-channel spectrograms for speech processing applications using deep learning methods

T. Arias-Vergara[1,2,3] · P. Klumpp[2] · J. C. Vasquez-Correa[1,2] · E. Nöth[2] · J. R. Orozco-Arroyave[1,2] · M. Schuster[3]

Abstract
Time–frequency representations of the speech signals provide dynamic information about how the frequency component changes with time. In order to process this information, deep learning models with convolution layers can be used to obtain feature maps. In many speech processing applications, the time–frequency representations are obtained by applying the short-time Fourier transform and using single-channel input tensors to feed the models. However, this may limit the potential of convolutional networks to learn different representations of the audio signal. In this paper, we propose a methodology to combine three different time–frequency representations of the signals by computing continuous wavelet transform, Mel-spectrograms, and Gammatone spectrograms and combining then into 3D-channel spectrograms to analyze speech in two different applications: (1) automatic detection of speech deficits in cochlear implant users and (2) phoneme class recognition to extract phone-attribute features. For this, two different deep learning-based models are considered: convolutional neural networks and recurrent neural networks with convolution layers.

Keywords Speech processing · Multi-channel spectrograms · Cochlear implants · Phoneme recognition

1 Introduction

In speech and audio processing applications, the data are commonly processed by computing compressed representations that may not capture the dynamic information of the signals. In the recent years, there has been an increasing number of works considering deep learning methods for speech and audio analysis such as convolutional neural networks (CNNs) and recurrent neural networks (RNN), among others [1]. Particularly for CNNs, audio data are processed by feeding the convolution layers with time–frequency representations (spectrograms) of the signals providing information about how the energy distributed in the frequency domain changes with time. After the convolution operation, the resulting feature maps contain low- and high-level features representing the acoustic information of the signals. Many works have shown the advantages of using CNNs and spectrograms in different speech processing applications such as automatic detection of disordered speech [2–4], acoustic models for automatic speech recognition systems [5, 6], and emotion detection [7], among others. These studies, however, consider single-channel spectrograms to obtain the feature maps, e.g., the short-time Fourier transforms (STFT) are applied to the audio signal and the resulting spectrogram is used as an input to the model. However, using only one representation may limit the potential of CNNs to learn more complex representations from the signals. One way to overcome this limitation is to use multiple spectrograms of each audio signal as input data to the CNN. For instance, multi-channel spectrograms were considered for audio source separation in [8, 9]. In those studies, audio recordings were captured with multiple microphones; thus, the multi-channel spectrograms are extracted from the same signal recorded with different microphones. A similar approach was presented in [10]. In this case, far-field automatic speech recognition is performed considering

---

✉ T. Arias-Vergara
  tomas.ariasvergara@lmu.de

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

3D-channel spectrograms, i.e., three different microphones were used to capture the speech signals. The main limitation of this approach is that it requires more than one microphone to obtain multiple spectrograms, which is not always possible (or necessary) in other applications, e.g., automatic detection of pathological speech. Multi-channel spectrograms can be also obtained from signals recorded with one channel. For instance, in [11] a methodology was presented to enhance noisy audio signals using complex spectrograms and CNNs. In that work, the real and imaginary part of the STFT is computed to form a 2D-channel spectrogram, which is then processed by the convolution layers; thus, the amplitude and phase information of the signal are considered to extract the feature maps.

In this study, we propose to combine Mel-spectrograms, Gammatone spectrograms (Cochleagrams), and continuous wavelet transform (CWT) to form multi-channel spectrograms. The proposed approach is then evaluated in two speech processing applications: automatic detection of disordered speech of cochlear implant (CI) users and phoneme class recognition to extract phone-attribute features. In our previous work [12], we showed that combining at least two different time–frequency representations of the signals can improve the automatic detection of speech deficits in CI users by training a bi-class CNN to differentiate between speech signals from CI users and healthy control (HC) speakers. This paper extends the use of multi-channel spectrograms to phoneme recognition using recurrent neural networks with convolutional layers (CRNN). For both, the CNN and CRNN, the first channel is the Mel-spectrogram, the second channel is the Cochleagram, and the third channel is the CWT of the speech signal. The Mel scale is inspired by findings of how humans perceive speech, which makes it suitable to represent the acoustic information of the sounds produced during speech. Cochleagrams are obtained with a Gammatone filter bank, which is based on the cochlear model proposed in [13], which 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). Both Mel and Gammatone spectrograms are computed based on the STFT whose time and frequency resolutions are determined by the size of the analysis window and the time-shift. A small window size can improve tie localization while resulting in poorer frequency resolution. Conversely, the larger we make the size of the window the more we will know about the frequency value and less about the time. Thus, the CWT is considered in this study to overcome this problem. The wavelet transform uses variations in a base function (called wavelet) highly localized in time. Each variation has a different scale, which allows to have high-frequency resolution for small-frequency values at the cost of low time resolution. At the same time, the CWT allows to have high time resolution at the cost of low-frequency resolution for high-frequency values. Our main hypothesis is that using

the spectrograms as a 3D-channel input will allow the CNN to complement the information from the two time–frequency representations. The rest of the paper is organized as follows: Sect. 2 describes the time–frequency analysis performed and the model architectures considered in this study. Section 3 describes the two applications for multi-channel spectrograms considered in this study. The data, preprocessing steps, and the training of the models (for each application) are also described in this section. Section 4 describes the experimental setup and the results obtained for each application. Finally, the conclusions derived from this work are presented in Sect. 5.

2 Methods

2.1 Time–frequency analysis

2.1.1 Mel/gammatone filterbanks

The STFT is the most commonly used time–frequency representation in speech and audio processing applications due to its simplicity and low computational cost. Alternatively, time–frequency representations can be obtained by applying a set of bandpass filters in the Mel scale (for Mel-spectrograms) or in the equivalent rectangular bandwidth (ERB) scale (for Cochleagrams) [14]. The log-Mel-spectrum is computed in three steps: First, the signal X is framed into short-time windows, i.e., $X = \{x_1, x_2, \ldots, x_T\}$ where T is the Tth speech frame. In this work, the size of the windows is 40 ms, which are extracted every 10 ms. In the next step, Hamming windows are applied to the framed signal in order to compute the STFT. In the last step, a set of 128 triangular filters in the Mel scale is applied and the logarithm of the resulting signal is computed in order to obtain the Mel-spectrum. Frequencies in Hz can be converted to Mel scale as:

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

The steps to obtain the Cochleagram are similar to the Mel-spectrum; however, it consists of bandpass filters in the ERB scale and the shape is obtained as the multiplication of sine and gamma functions. The Gammatone filter bank is defined in the time domain by Eq. 2 as:

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

where f_c is the filter's center frequency in Hz, ϕ 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 Gammatone filters are implemented following the procedure described in [15]. The number of filters used for both Mel-scale and Gammatone based features is $n = 128$. Figure 1 shows the triangular and Gammatone filter banks considered in this study.
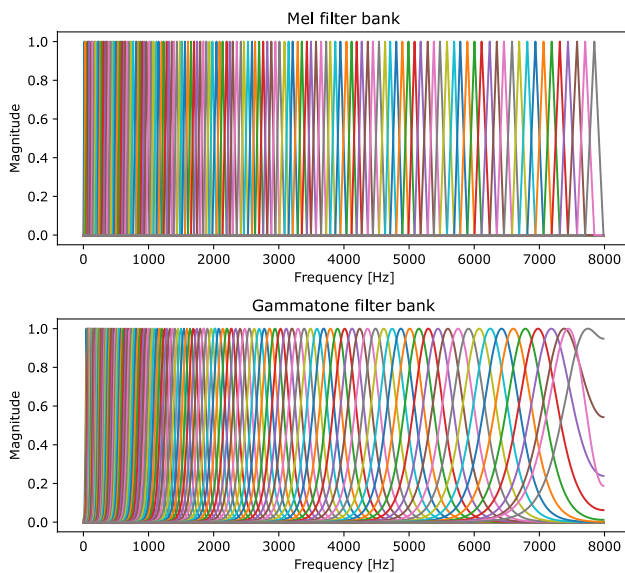
Fig. 1 Set 128 triangular filters in the Mel scale and 128 Gammatone filters in the ERB scale are applied to the STFT in order to obtain the Mel-spectrum and the Cochleagram, respectively

2.1.2 Continuous wavelet transform

Contrary to the STFT, the time and frequency resolutions of the CWT are not determined by the size of the analysis window and the time-shift. Instead, the CWT considers a base function called wavelet in order to decompose the speech signal. This procedure is performed by convolving the signal with shifted and compressed versions of the wavelet. Formally, the CWT is defined as

$$\text{CWT}(u, s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} x(t) \psi^* \left(\frac{t - u}{s} \right) dt \tag{3}$$

where $x(t)$ is the speech signal, u and s are the shift and scale parameters, respectively, and ψ is the mother wavelet (base function), which in this study is the Morlet wavelet. Figure 2 shows the Mel-spectrum, Cochleagram, and resulting CWT of a speech signal. The Mel-spectrum and the Cochleagram are obtained after applying filter banks to the STFT of the speech signal. The output of the CWT consists of a two-dimensional time-scale representation of the speech signal. In our case, the number of scales used ranges from 1 to 128, in order to match the dimensions of the Mel-spectrum and the Cochleagram in the frequency dimension.

2.2 Model architectures

Two different models are used to test the suitability of multi-channel spectrograms for speech processing applications. The first method consists of a CNN for automatic detection of disordered speech. The convolution layer in a CNN acts like a filter bank, which allows to capture high- and low-level features from the spectrograms [16, 17]. The second method consists of a convolutional recurrent neural network with gated recurrent units (CGRU) for phoneme recognition. The main advantage of using recurrent networks is their ability to learn contextual information from speech sequences [18], which makes them suitable for speech recognition applications. For both methods, the input tensor of the convolutional layers consists of the Mel-spectrogram in one channel, the Cochleagram in the second channel, and the CWT in the
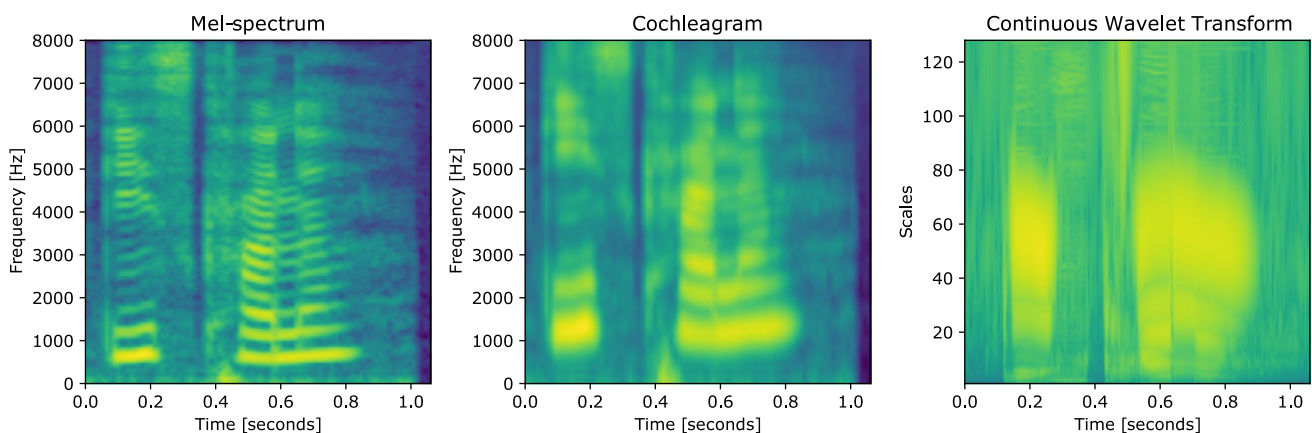

Fig. 2 Mel-spectrum, Cochleagram, and CWT of a speech signal. The Mel-spectrum is obtained after applying a set of triangular filter bank (in the Mel scale) to the STFT of the speech signal. The Cochleagram is obtained after applying a Gammatone filter bank (in the ERB scale) to the STFT. The CWT is obtained after convolving a Morlet wavelet (with a linear scale from 1 to 128) with the speech signal
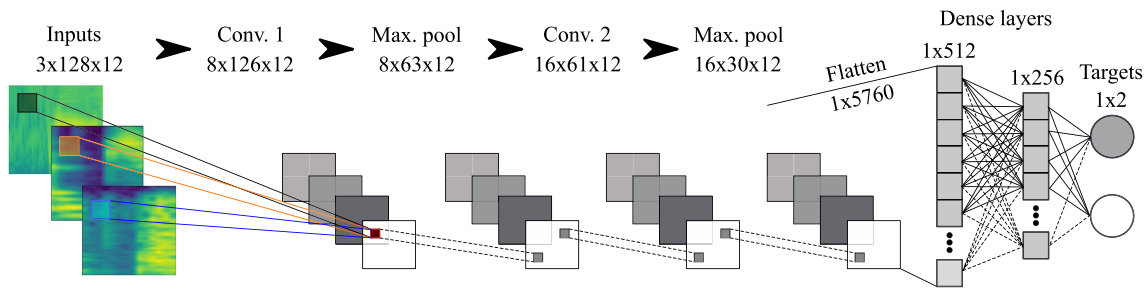
Fig. 3 Architecture of the CNN implemented in this study. The size of the kernel in the convolutional (Conv. *i*) and pooling layers (Max. pool) is 1×3 and 1×2, respectively
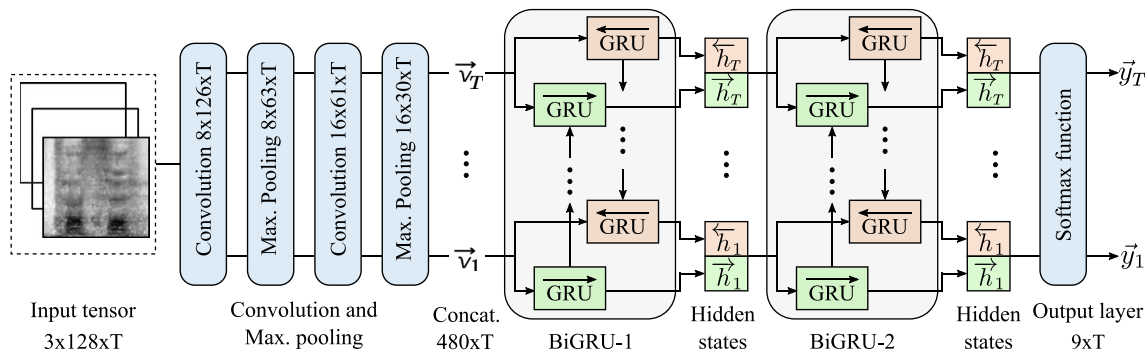

Fig. 4 CGRU architecture considered in this work. The input sequences are 3D-channel inputs formed with Mel-spectrograms, Cochleagrams, and CWT with Morlet wavelets. Convolution is performed only on the frequency axis to keep the time information. The resulting feature maps are then feed into a 2-stacked bidirectional GRU. A softmax function is then used to predict the phoneme label for every speech segment in the input signal

third channel. The framework PyTorch [19] is considered to implement the proposed architectures. From the documentation, it can be observed that the output of the convolutional layer for an input signal is described as:

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

where *Bs* is the batch size and *C* is the number of channels of the input tensor ($C = 3$). The following subsections describe the architectures implemented in this study.

### 2.2.1 Convolutional neural network

There are no standard guidelines to determine the optimal architecture of a CNN. Commonly, the best configuration is chosen experimentally based on performance evaluation. Instead of trying different architectures, we test the multi-channel spectrograms by adapting the LeNet-5 convolutional network [20]. The configuration of our network consists of two convolution layers with rectifier linear (ReLU) activation functions, two max-pooling layers, dropout to regularize the weights, and two fully connected hidden layers

followed by the output layer to make the final decision using a softmax activation function. The CNN is trained using the Adam optimization algorithm [21] with a learning rate of $\eta = 10^{-4}$. The cross–entropy between the training labels *y* and the model predictions $\hat{y}$ is used as the loss function. The size of the kernel in the convolution layers is $k_c = 1 \times 3$. For the pooling layers, the kernel's size is $k_p = 1 \times 2$. Note that the convolution and pooling operations are performed only in one dimension (frequency/scale). The reason is that we want to keep as much information from the time dimension as possible. Figure 3 summarizes the configuration of the network used in this work. The number of output channels in the first and second convolution layers is 8 and 16, respectively. The size of the second layer is twice the size of the first convolution layer in order to allow the network to extract high-level features from the speech signals [17].

2.2.2 Recurrent neural network with convolution layers

The architecture of the CGRU is summarized in Fig. 4. The multi-channel spectrograms are processed by two convolution layers with ReLU activation functions, two max-pooling layers, and dropout to regularize the weights. The

Fig. 5 Onset and offset transitions extracted from a speech recordings. The transitions consist of speech segments of 160 ms containing voiceless and voiced segments
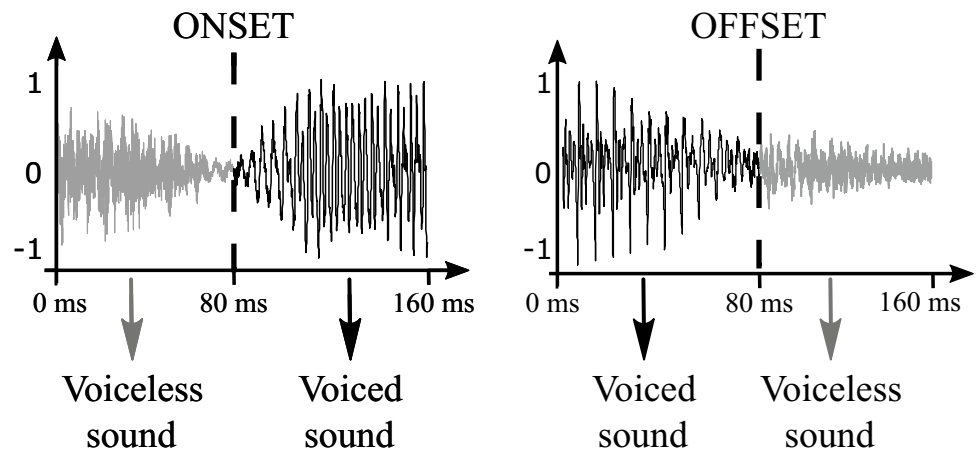
convolution and max-pooling operations are performed only on the frequency axis of the 2D-channel spectrograms in order to keep a one-to-one relation between the length of the input (speech sequences) and the output (phoneme prediction). The size of the kernel in the convolutional (Conv. i) and pooling layers (Max. pool) is 1×3 and 1×2, respectively. After convolution, the resulting feature maps are concatenated to form the sequence of feature vectors $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T\}$, where T is the total number of frames. The sequence $\mathbf{V}$ is then processed by two bidirectional recurrent layers (BiGRU-1 and BiGRU-2) with shared weights on each time frame t. Thus, for every input data $\mathbf{v}_t$ in the sequence, the network has sequential information about the data points before $(\mathbf{v}_1, \ldots, \mathbf{v}_{t-2}, \mathbf{v}_{t-1})$ and after $(\mathbf{v}_{t+1}, \mathbf{v}_{t+2}, \ldots, \mathbf{v}_T)$ [22]. A softmax activation function is used to compute the sequence of phoneme posterior probabilities $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$. Bidirectional recurrent nets are used in this work because they have shown better results than standard GRUs in similar speech processing tasks [23, 24].

Similar to the CNN, the CGRU is trained using the same optimization algorithm, learning rate, and loss function. Note that some phonemes are more frequently produced than others. For instance, the number of vowels is higher than the number of nasal sounds in the database. As a result, the performance of the system to detect the phoneme classes that are underrepresented is lower than the phonemes that are more commonly produced. Thus, class weights are introduced into the loss function, which is described as:[1]

$$\text{loss}(p, \text{class}) = w[l]\left(-p[l] + \log\left(\sum_j \exp(p[j])\right)\right) \quad (5)$$

where p are the posterior probabilities of the sequences obtained from the output layer $y = \{y_1, y_2, \ldots, y_T\}$, l are the target labels, and w are the class weights.

3 Applications

### 3.1 Automatic detection of disordered speech in CI users

Cochlear implants (CI) are the most suitable devices for severe and profound deafness when hearing aids do not improve sufficiently speech perception. However, CI users often present altered speech production and limited understanding even after hearing rehabilitation. People suffering from severe to profound deafness may experience different speech disorders such as decreased intelligibility and changes in terms of articulation [25]. Acoustic analysis is performed in order to detect articulatory problems in the speech of CI users by detecting the voiceless-to-voiced (Onset) and voiced-to-voiceless (Offset) transitions, which are considered to model the difficulties of the CI users to start/stop the movement of the vocal folds [26, 27]. 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 [28]. 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 (Fig. 5).

3.1.1 Data: CI speech

Standardized speech recordings of 107 CI users (56 male) and 94 HC (46 male) are considered for the experiments. All

---

Fig. 6 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
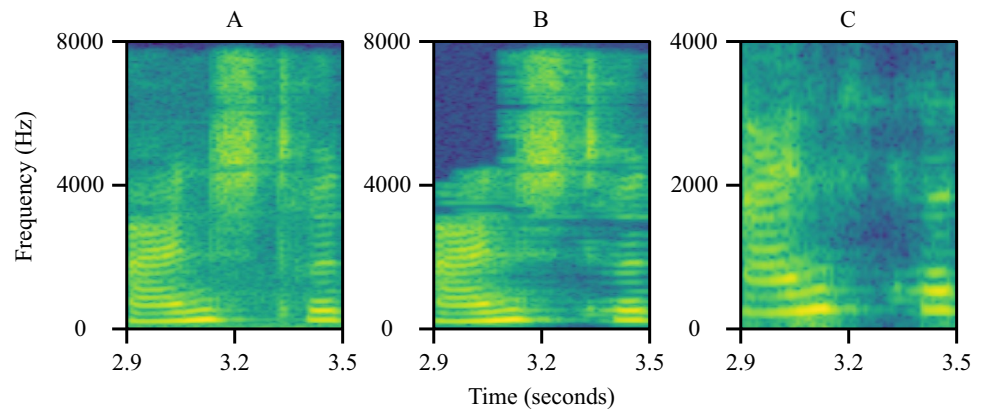
of them are German native speakers. The speech signals of the CI users were recorded at the clinic of the Ludwig-Maximilians University in Munich (LMU). The recordings of the HC speakers were extracted from the PhonDat 1 (PD1) corpus from the Bavarian Archive For Speech Signals (BAS), which is freely available for European academic users.[2] The speech recordings include the reading of *Der Nordwind und die Sonne* (*The North Wind and the Sun*) text.

3.1.2 Preprocessing

Note that some of the recordings from the HC speakers were collected in different acoustic conditions than the speakers recorded in the clinic; thus, noise reduction and compression techniques are applied to the speech signals in order to reduce the effect of the channel in the recordings.

Noise reduction Background noise is reduced based on the spectral gating algorithm implemented in the SoX codec.[3] The core idea of the algorithm is to attenuate the speech segments in the signal with spectral energy below certain thresholds, which are obtained by computing the mean power on each frequency band from the STFT of a noise profile extracted from a silence region of the speech signal.

Compression After noise reduction, the GSM full-rate compression technique is considered to normalize the channel conditions of the recordings [29]. First, the denoised signals are down-sampled to 8 kHz and the resolution is lowered down to 13 bits, with a compression factor of 8. Next, a bandpass filter between 200 Hz and 3.4 kHz is applied in order to meet the specifications of a GSM transmission network. Figure 6 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 recording of one of the healthy speakers in the database.

3.1.3 Training of the CNN

Onset and offset transitions are extracted from the speech recordings in order to train the CNN described in Sect. 2.2.1. A tenfold cross-validation strategy is considered in order to train and test the models. The performance of the CNN is measured by means of precision, recall, and F1-score. Precision measures the proportion of predicted speech segments (onset/offset transition) that are correctly classified. Recall measures the proportion of actual speech segments that are correctly classified. The F1-score measures the performance of the CNN to classify all speech segments, which reaches its best value at 1 and worst score at 0. These three measures are computed as in [30].

3.2 Phone-attribute features

Previous work has shown the suitability of phone-attribute features to evaluate articulation precision in people learning a second language [31] as well as to evaluate speech problems in patients affected by different medical conditions such as Parkinson's disease [32] and hearing loss [33]. In this work, phone-attribute features are computed using the CGRU described in Sect. 2.2.2 which converts speech sequence $X = \{x_1, x_2, \ldots, x_T\}$ into a sequence of posterior probabilities $y = \{y_1, y_2, \ldots, y_T\}$, where T is the number of frames extracted from the speech signal. The speech sequences consist of Mel-spectrograms, Cochleagrams, and CWT. The vector of phone-attribute features $y_n = \{y_n^1, \ldots, y_n^k, \ldots, y_n^K\}$ consists of K phoneme probabilities (posteriors). The CGRU estimates the posterior y_n^k as the probability of occurrence of the kth phone-attribute feature. The main hypothesis is that normal speakers can produce phonemes correctly; thus, the posterior probabilities of occurrence of phonemes (phone-attribute features) are close to 1. On the other hand, if the model is tested with a speech

---

Table 1 Phone-attribute features considered in this study

Feature	Brief description
Stop	Refers to sounds produced by a total oral closure that is rapidly released
Trill	Phonemes produced by a vibration of the two articulators, caused by the current of air
Nasal	Refers to sounds produced by a lowered velum such that air can flow through the nasal cavity
Vowel	Sounds produced by the vibration of the vocal folds and the changes in the shape of the vocal tract
Silence	Regions of the signal with the absence of speech
Lateral	The air passes at the sides of the tongue which forms a central closure
Fricative	A constriction in the articulator causes a turbulence in the flowing air, producing a hissing sound
Sibilants	Similar to fricatives, however, these phonemes have more acoustic energy in higher frequency bands
Approximants	Similar to fricatives, however, the constriction is not narrow enough to cause turbulences in the air flow

Table 2 Information on the partitions

	Test	Val	Train
Male speakers	14	19	275
Female speakers	15	3	260
Hours of speech	1.5	1.5	26
Number of recordings	585	584	10,845

Test Test set, *Val* validation set, *Train* train set

signal from a speaker with pronunciation problems, then the posterior probability will be lower compared with respect to the normal speaker. In this paper, the phone-attribute feature are computed for nine phoneme classes (including "silence") which are grouped according to the standard German language system. A short description of the phone-attribute features is presented in Table 1.

3.2.1 Data: Verbmobil

The Verbmobil corpus consists of speech recordings from 586 German native speakers (308 male, 278 female). The database contains about 29 hours of dialogues with their corresponding phonetic transcriptions. The data were captured in controlled acoustic conditions with a close-talk microphone at a sampling frequency of 16 kHz and a resolution of 16-bit. The age of the speakers ranges from 20 up to 40 years [34].

3.2.2 Training of the CGRU

Chunks of data of 1 s are extracted from the speech recordings in order to train the CGRU, i.e., the input data consist of speech sequences with a fixed length. Each sequence is then time-aliment with their corresponding phonetic transcription; thus, each time-frame is labeled according to one of the nine phoneme classes described in Table 1. The input tensors and their corresponding target labels are then used to train the CGRU for phone-attribute feature extraction.

Table 3 Classification results between speech segments of CI users and HC speakers

Model	Channel	Prec	Rec	*F*1
Onset	CWT	0.78	0.77	0.77
	Mel-spectrum	0.82	0.80	0.81
	Cochleagram	0.81	0.81	0.81
	3D Spectrum	0.84	0.82	0.82
Offset	CWT	0.80	0.78	0.78
	Mel-spectrum	0.84	0.82	0.82
	Cochleagram	0.84	0.80	0.81
	3D Spectrum	0.86	0.83	0.84

Prec Precision, *Rec* recall, *F*1 *F*1 score, *CWT* continuous wavelet transform, *3D-spectrum* mel-spectrum, Cochleagram, and CWT

Table 2 shows the information about the train, validation, and test sets considered in this study. The performance of the model is evaluated by means of the precision (the ability of the CGRU not to label as positive a sample that is negative), recall (the ability of the CGRU to correctly label the phonemes classes), and F1 score (weighted harmonic mean of the precision and recall) [30].

4 Experiments and results

4.1 Multi-channel spectrograms with CNN

Table 3 shows the results obtained when the CNN is trained to classify speech segments (onset/offset transitions) from CI users and HC speakers. The highest classification performance is obtained with three-channel spectrograms extracted from the offset transitions ($F1 = 0.84$). Note also that the results obtained with the Mel-spectrum and the Cochleagram are similar for both onset and offset transitions. This can be explained considering that these time–frequency representations are obtained from the same transformation, i.e., the STFT. Furthermore, the lowest performance was

Table 4 Performance of the CGRU for phoneme class recognition

Phoneme class	CWT			Mel-spectrum			Cochleagram			3D spectrum		
	Prec	Rec	*F*1	Prec	Rec	*F*1	Prec	Rec	*F*1	Prec	Rec	*F*1
Silence	0.89	0.86	0.87	0.91	0.86	0.88	0.90	0.87	0.88	0.90	0.86	0.88
Stop	0.81	0.80	0.80	0.83	0.82	0.82	0.83	0.82	0.82	0.84	0.80	0.82
Nasal	0.80	0.83	0.82	0.82	0.85	0.83	0.81	0.85	0.83	0.81	0.85	0.83
Trill	0.00	0.00	0.00	0.80	0.67	0.73	0.76	0.68	0.72	0.78	0.69	0.73
Fricative	0.79	0.80	0.79	0.84	0.82	0.83	0.83	0.83	0.83	0.82	0.83	0.82
Sibilant	0.84	0.89	0.86	0.86	0.90	0.88	0.85	0.90	0.88	0.85	0.90	0.88
Approximant	0.81	0.79	0.80	0.79	0.84	0.81	0.80	0.82	0.81	0.84	0.78	0.81
Lateral	0.69	0.60	0.64	0.71	0.68	0.70	0.73	0.68	0.70	0.68	0.70	0.69
Vowel	0.82	0.86	0.84	0.84	0.87	0.85	0.85	0.86	0.85	0.84	0.86	0.85

Prec Precision, *Rec* recall, *F*1 *F*1 score

obtained when only the CWT is considered as input to the CNN (F_1-Onset = 0.77; F_1-Offset = 0.78).

4.2 Multi-channel spectrograms with CGRU

Table 4 shows the results obtained for the automatic detection of the phoneme classes described in Table 1. On the one hand, it can be observed that the performance of the CGRU is similar when is trained with Mel-spectrograms, Cochleagrams, and 3D-channel spectrograms; thus, the contribution of three channels is not decisive enough to improve the phoneme class recognition. On the other hand, the performance of the CGRU trained with the CWT is lower than for Mel-spectrum and Cochleagram in all classes. Particularly, it can be observed that it was not possible to detect any phoneme from the class "Trills."

5 Conclusion

In this paper, Mel-spectrograms, Cochleagrams, and CWT are combined to form three-channel spectrograms. Two different applications were considered: (1) automatic detection of disordered speech of CI users and (2) phoneme class recognition to extract phone-attribute features. In the first application, speech signals of CI users and HC were considered to train a CNN to perform binary classification. The CNN was trained considering Mel-spectrograms, Cochleagrams, CWT, and the combination of the three representations. Additionally, onset and offset transitions are extracted from the speech signals in order to perform acoustic analysis to evaluate the articulatory precision of the speakers. According to the results, the highest performance was achieved when the CNN was trained with the 3D-channel spectrograms extracted from the offset transitions. In the second application, a CGRU was trained to automatically recognize phonemes grouped in seven different classes. The model was trained with recordings of normal speakers, i.e., people

without any speech disorder or neurological disease. From the results, it was observed that the contribution of the multi-channel spectrograms was not decisive enough to improve the recognition of phoneme classes. One hypothesis is that the way the spectrograms are combined does not provide sufficient information for the network to learn a proper representation of the phoneme classes; thus, future work should focus on different configurations of the network or include different time–frequency representations. Furthermore, the models should be trained and tested with noisy signals in order to test the robustness of the classifiers for speech signals captured in non-controlled acoustic conditions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Purwins H, Li B, Virtanen T, Schlüter J, Chang S, Sainath T (2019) Deep learning for audio signal processing. IEEE J Sel Top Signal Process 13(2):206–219

2. Vásquez-Correa JC, Orozco-Arroyave JR, Nöth E (2017) 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, pp 314–318

3. Wu H, Soraghan J, Lowit A, Di Caterina G (2018) A deep learning method for pathological voice detection using convolutional deep belief networks. In: Proceedings of the nineteenth annual conference of the international speech communication association, pp 446–450

4. Alhussein M, Muhammad G (2018) Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access 6:41034–41041

5. Abdel-Hamid O, Mohamed A, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(10):1533–1545

6. Han K, He Y, Bagchi D, Fosler-Lussier E, Wang D (2015) Deep neural network based spectral feature mapping for robust speech recognition. In: Sixteenth annual conference of the international speech communication association, pp 2484–2488

7. Weißkirchen N, Bock R, Wendemuth A (2017) Recognition of emotional speech with convolutional neural networks by means of spectral estimates. In: 2017 seventh international conference on affective computing and intelligent interaction workshops and demos (ACIIW), pp 50–55

8. Adavanne S, Politis A, Virtanen T (2018) Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features. In: 2018 international joint conference on neural networks (IJCNN), pp 1–7

9. Xu K, Feng D, Mi H, Zhu B, Wang D, Zhang L, Cai H, Liu S (2018) Mixup-based acoustic scene classification using multi-channel convolutional neural network. In: Pacific Rim conference on multimedia, pp 14–23

10. Ganapathy S, Peddinti V (2018) 3-D CNN models for far-field multi-channel speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5499–5503

11. Fu S, Hu T, Tsao Y, Lu X (2017) Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: 2017 IEEE 27th international workshop on machine learning for signal processing (MLSP), pp 1–6

12. Arias-Vergara T, Vasquez-Correa JC, Gollwitzer S, Orozco-Arroyave JR, Schuster M, Nöth E (2019) Multi-channel convolutional neural networks for automatic detection of speech deficits in cochlear implant users. In: Iberoamerican congress on pattern recognition, pp 679–687

13. Patterson RD, Robinson K, Holdsworth J, McKeown D, Zhang C, Allerhand M (1992) Complex sounds and auditory images. Elsevier, Amsterdam, pp 429–446

14. Virtanen T, Vincent E, Gannot S (2018) Time-frequency processing-spectral properties. In: Audio source separation and speech enhancement, pp 15–29

15. Slaney M, et al (1993) An efficient implementation of the Patterson–Holdsworth auditory filter bank. Apple Computer, Perception Group, Technical Report 35(8)

16. Latif S, Rana R, Khalifa S, Jurdak R, Qadir J, Schuller B (2020) Deep representation learning in speech processing: challenges, recent advances, and future trends. arXiv:2001.00378

17. Palaz D, Collobert RN, et al (2015) Analysis of CNN-based speech recognition system using raw speech as input. Technical Reports, Idiap

18. Graves A (2012) Supervised sequence labelling with recurrent neural networks, vol 385

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

20. LeCun Y, Bottou Ln, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

21. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: International conference on learning representation (ICLR)

22. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681

23. Cernak M, Tong S (2018) Nasal speech sounds detection using connectionist temporal classification. In: IEEE, pp 5574–5578

24. Vásquez-Correa JC, Klumpp P, Orozco-Arroyave JR, Nöth E (2019) Phonet: a tool based on gated recurrent neural networks to extract phonological posteriors from speech, pp 549–553

25. Hudgins CV, Numbers FC (1942) An investigation of the intelligibility of the speech of the deaf. In: Genetic psychology monographs

26. Arias-Vergara T, Gollwitzer S, Orozco-Arroyave JR, Vasquez-Correa JC, Nöth E, Högerle C, Schuster M (2019) Speech differences between CI users with pre-and postlingual onset of deafness detected by speech processing methods on voiceless to voice transitions. Laryngo-Rhino-Otologie 98(S02):11435

27. Arias-Vergara T, Orozco-Arroyave JR, Gollwitzer S, Schuster M, Nöth E (2019) Consonant-to-vowel/vowel-to-consonant transitions to analyze the speech of cochlear implant users. In: International conference on text, speech, and dialogue, pp 299–306

28. Orozco-Arroyave JR (2016) Analysis of speech of people with Parkinson's disease. Logos Verlag, Berlin

29. Huerta JM, Stern RM (1998) Speech recognition from GSM codec parameters. In: Fifth international conference on spoken language processing, pp 1–4

30. Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

31. Arora V, Lahiri A, Reetz H (2017) Phonological feature based mispronunciation detection and diagnosis using multi-task DNNs and active learning

32. Garcia-Ospina N, Arias-Vergara T, Vásquez-Correa JC, Orozco-Arroyave JR, Cernak M, Nöth E (2018) Phonological I-vectors to detect Parkinson's disease. In: International conference on text, speech, and dialogue, pp 462–470

33. Arias-Vergara T, Orozco-Arroyave JR, Cernak M, Gollwitzer S, Schuster M, Nöth E (2019) Phone-attribute posteriors to evaluate the speech of cochlear implant users. In: Proceedings of the 20th annual conference of the international speech communication association, pp 3108–3112

34. Wahlster W (2013) Verbmobil: foundations of speech-to-speech translation. Springer, Berlin