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

Phonet: a Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech

J. C. Vásquez-Correa[1,2]*, P. Klumpp*[1]*, J. R. Orozco-Arroyave*[1,2]*, and E. Nöth*[1]

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]Faculty of Engineering, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
juan.vasquez@fau.de

Abstract

There are a lot of features that can be extracted from speech signals for different applications such as automatic speech recognition or speaker verification. However, for pathological speech processing there is a need to extract features about the presence of the disease or the state of the patients that are comprehensible for clinical experts. Phonological posteriors are a group of features that can be interpretable by the clinicians and at the same time carry suitable information about the patient's speech. This paper presents a tool to extract phonological posteriors directly from speech signals. The proposed method consists of a bank of parallel bidirectional recurrent neural networks to estimate the posterior probabilities of the occurrence of different phonological classes. The proposed models are able to detect the phonological classes with accuracies over 90%. In addition, the trained models are available to be used by the research community interested in the topic.
Index Terms: Phonological posteriors, Gated recurrent units, Recurrent neural network, Pathological speech

1. Introduction

Different groups of features are commonly used in speech processing applications. For instance, Mel frequency cepstral coefficients (MFCCs) are widely used in automatic speech recognition (ASR) [1], or the perceptual linear predictive coefficients (PLPs), which have shown to be suitable for speaker identification [2]. Recently, deep learning methods have also been considered to extract features from speech for specific tasks [3, 4]. However, for pathological speech processing, only a small subset of basic features are commonly used, e.g., fundamental frequency, jitter, shimmer, or formant frequencies [5, 6]. More complex feature sets like MFCCs, PLPs, or embeddings from neural networks are rarely used for the medical community due to their lack of interpretability. However, these high-dimensional feature vectors contain a great amount of information about the state of the patients that can be exploited by clinicians. Thus, it is important to provide clinically interpretable features that at the same time carry suitable information about the health state of the patients.

Phonological features can be more comprehensible for clinicians than the traditional high-dimensional acoustic features used in speech processing. Phonological features are represented by a vector with explainable information about the mode and manner of articulation of the speaker. These dimensions are commonly understood by clinicians since the features are related specifically with the movements of the articulators in the vocal tract. This study aims to map high-dimensional feature vectors into explainable feature vectors called "phonological posteriors" that can be comprehensible for the medical community. The use of phonological posteriors has been considered for several applications related to pathological speech, including assessment of dysarthric speech [7, 8], or evaluation of progressive apraxia of speech [9]. In [10] the authors combined information from speech, facial expression, and electroencephalography (EEG) to detect phonological categories such as consonant, nasal, bilabial, and others. The model was trained with utterances from isolated phonemes, syllables, and English words. The authors reported accuracies over 90% in the detection of consonants using a deep-belief network. In [11] the authors presented a toolkit to estimate 15 phonological posteriors in English language. The authors considered a parallel bank of fully connected networks to recognize each phonological class. Accuracies over 96% were reported for detecting the phonological classes, including nasal, strident, and vocalic. In [7] the authors detect phonological classes using a model based on recurrent neural networks (RNNs) with long short-term memory units (LSTM) trained with the TIMIT corpus. Phonological features were used to evaluate the articulation quality of dysarthric speakers. The authors considered 15 phonological features, which are detected with accuracies over 90%. These phonological features extracted from utterances of dysartrhic speakers were correlated with a perceptual rating of the articulatory precision of the subjects. The authors reported Pearson's correlations of up to 0.79 between the phonological features and the perceptual score assigned by speech therapists. The phonological features trained in [11] were used to predict the dysarthria level of patients with Parkinson's disease [8]. The authors correlated the phonological posteriors estimated for utterances of 50 patients and 50 healthy subjects with a subjective evaluation performed by speech therapists following a modified version of the Frenchay dysarthria assessment (m-FDA) scale [12]. The authors reported a moderate Spearman's correlation (0.56) even though the phonological features were trained with US English utterances and the Parkinson's patients were Colombian Spanish native speakers. Better results could be obtained if the phonological features were trained using the same language as that of the speakers to be evaluated.

Although the success on the use of phonological features to characterize pathological speech, there is a lack of models available for the research community that can be used and adapted for different pathological speech applications. The availability of models is even more scarce for languages different to English. We propose a model to estimate phonological posteriors based on bidirectional RNNs with gated recurrent units (GRUs). The model is trained with Spanish language utterances to test the reliability of the phonological analysis in a language different to English. The trained models using Keras [13] are available online [1] as a toolkit to be used by the research community interested in pathological speech assessment.

[1]https://github.com/jcvasquezc/phonet

http://dx.doi.org/10.21437/Interspeech.2019-1405

2. Materials and Methods

2.1. Phonological Posteriors

The phonetic alphabet for Spanish includes 24 different phonemes, represented by 5 vowels and 19 consonants [14]. These phonemes can be grouped into phonological classes based on the mode and manner of articulation of the sounds. Tables 1 and 2 show the distribution of the Spanish phonemes into the phonological classes for vowels and consonants, respectively. The notation of the phonemes is based on the international phonetic alphabet (IPA).

Table 1: *Distribution of the Spanish vowel phonemes.*

	Front	**Central**	**Back**
High	/i/		/u/
Mid	/e/		/o/
Low		/a/	

Table 2: *Distribution of the Spanish consonant phonemes.* ***Cont.*** *Continuant,* ***Alv.*** *Alveolar,* ***Pal.*** *Palatal*

	Labial		**Dental**		**Alv.**	**Pal.**	**Velar**	
Nasal	/m/				/n/	/ɲ/		
Stop	/p/	/b/	/t/	/d/		/tʃ/	/k/	/g/
Cont.	/f/		/θ/		/s/	/ʝ/	/x/	
Lateral					/l/	/ʎ/		
Flap					/ɾ/			
Trill					/r/			

Some conventions were considered in this study to extract different phonological classes based on the Spanish phonemes: (1) the phoneme /θ/ was not considered because it is only used in Spanish from central Spain. (2) The phoneme /ʝ/ from the word "cayado" and the phoneme /ʎ/ from the word "callado" were grouped together since they are pronounced similarly in many Latin American countries [15]. (3) The phoneme /n/ from the word /cana/ and the phoneme /ɲ/ from the word "caña" were also grouped together because they belong to the same phonological categories considered in this study. Based on the defined conventions, we have a phonetic alphabet with 21 phonemes to train the proposed models. Those phonemes are distributed into 18 phonological classes defined according to Table 3 based on the movement of different articulators in the vocal tract. The "pauses" are considered as a separate phonological class.

The phonological posteriors will be the conditional posterior probability of a speech frame to belong to one or more phonological classes. The phonological posteriors will be computed with a bank of parallel RNNs, which estimate the probability of occurrence of a specific phonological class. Only very few phonological classes are active during a short term signal, which results in a sparse vector representation [11]. With the aim to complement the estimation of phonological posteriors, we propose here an additional model to recognize the 21 phonemes grouped into the phonological classes.

2.2. Deep learning model

The proposed model to extract the phonological posteriors is shown in Figure 1. The speech signals are segmented into "chunks" of 0.5 seconds to be used as inputs to the neural network. Each "chunk" is windowed into frames of 25 ms to compute the feature sequence for the input layer of the neural network. The input features correspond to the log-energy of the signal distributed in 33 triangular filters separated according to the Mel scale. The feature sequences from the input are processed by two bidirectional GRU layers to model information from the past (backward) and future (forward) states of the sequence, simultaneously. The GRUs were proposed as a modification of the LSTMs, replacing the separate input and forget gates with a reset gate to control the input information to the network. GRUs and LSTMs have provided similar results for several tasks, including speech and language modeling [16]; however, the GRUs are faster to train and require less parameters [17], which make these units more suitable to be used when less training data is available. The output sequences of the second bidirectional GRU layer pass through a time distributed hidden dense layer, which keeps a one-to-one relation between the length of input and output sequences, i.e., it applies a fully connected dense layer with shared weights on each time-step, producing an output sequence of the same length as the input. The time distributed hidden layer is connected to the time distributed output layer with a softmax activation function, which produces the sequence of posterior probabilities for a phonological class associated to the feature sequence from the input.

Table 3: *Distribution of the different Spanish phonemes into phonological classes.*

Phonological class	**List of phonemes**
Vocalic	/a/, /e/, /i/, /o/, /u/
Consonantal	/b/, /tʃ/, /d/, /f/, /g/, /x/, /k/, /l/, /ʎ/, /m/, /n/, /p/, /ɾ/, /r/, /s/, /t/
Back	/a/, /o/, /u/
Anterior	/e/, /i/
Open	/a/, /e/, /o/
Close	/i/, /u/
Nasal	/m/, /n/
Stop	/p/, /b/, /t/, /k/, /g/, /tʃ/, /d/
Continuant	/f/, /b/, /tʃ/, /d/, /s/, /g/, /ʎ/, /x/
Lateral	/l/
Flap	/ɾ/
Trill	/r/
Voice	/a/, /e/, /i/, /o/, /u/, /b/, /d/, /l/, /m/, /n/, /r/, /g/, /ʎ/
Strident	/f/, /s/, /tʃ/
Labial	/m/, /p/, /b/, /f/
Dental	/t/, /d/
Velar	/k/, /g/, /x/
Pause	/sil/

The architecture from Figure 1 is used to train a bank of 19 neural networks: 18 corresponding to the phonological classes defined in Table 3 and one to recognize the 21 phonemes considered in this study. The different networks are trained with a weighted categorical cross-entropy loss function, defined according to Equation 1 to avoid the unbalance of the classes in the training process. The weight factors w_i for each class $i = \{1 \cdots C\}$ are defined based on the percentage of samples from the training set that belong to each class. The networks were trained using an Adam optimizer [18]. In addition, dropout and batch normalization layers were considered to improve the generalization of the proposed networks.

$$\mathcal{L} = -\sum_{i=1}^{C} w_i p_i \log(\widehat{p_i}) \tag{1}$$

3. Data

The training of the proposed models is performed with the CIEMPIESS corpus [19], which consists of 17 hours of FM

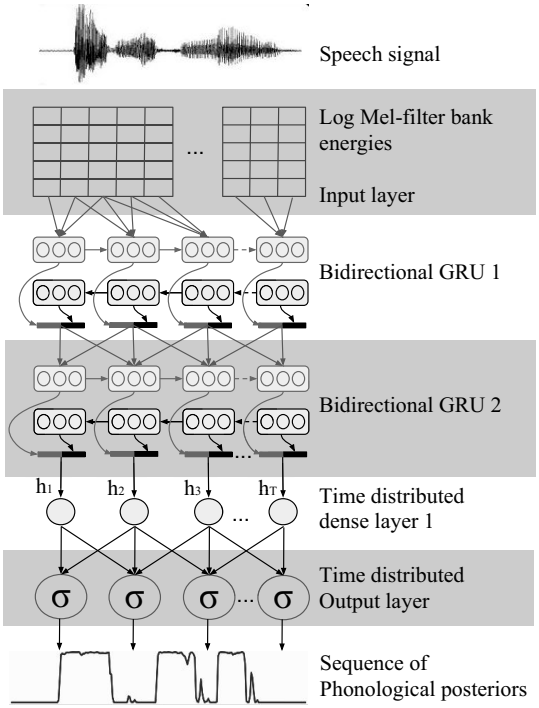

Figure 1: *Architecture of the proposed neural network to estimate the phonological posteriors from the speech.*

podcasts in Mexican Spanish. The database was designed to be used in ASR systems, and it was annotated at word level, considering all the phonemes of the Spanish language. The data consider only "clean" utterances i.e., those made by only one person, with no background noises, foreign accents, or music. The data contain 16717 audio files with a sampling frequency of 16 kHz and 16-bit resolution. 700 utterances from the entire corpus (speaker independent) were subtracted to be used as the test set of the experiments performed in this study. The complete corpus was forced-aligned using the *BAS CLARIN* web service[2] [20] based on the phonetic segmentation introduced in [21] for Spanish. The audio files and their corresponding transcriptions were uploaded to the server, which provides *Textgrid* files with the phonetic alignment for each utterance. The aligned phonemes were used as labels to train our models for phoneme recognition and for the estimation of the phonological posteriors.

4. Experiments and Results

4.1. Phonological Posteriors

The results of the proposed method to recognize the different phonological classes are shown in Table 4. The proposed model shows to be highly accurate to detect the different phonological classes. The unweighted average recall (UAR) ranges from 80.4% to 93.3%, depending on the phonological class. Note especially the high accuracy obtained for the *strident*, *nasal*, *Labial*, *Pauses*, and *anterior* phonological classes.

Figure 2 shows an example of the "posteriorgram" obtained for a speech signal of the Spanish sentence "mi casa tiene tres cuartos" (my house has three rooms). The figure shows the activation of the different phonological classes through time.

[2] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

Table 4: *Results of the recognition of the different phonological classes using the proposed approach.* ***UAR (%)****. Unweighted average recall,* ***Spec. (%)*** *Specificity,* ***Sens. (%)*** *Sensitivity.*

Phonological Class	**UAR**	**Spec.**	**Sens.**	**F-score**
Anterior	89.7	87.0	92.3	0.889
Back	88.2	87.5	88.9	0.882
Close	85.9	89.7	82.1	0.901
Consonantal	83.3	81.5	85.1	0.831
Continuant	85.7	84.0	87.4	0.861
Dental	85.8	87.6	84.0	0.898
Flap	80.4	86.3	74.4	0.895
Labial	89.3	86.5	92.0	0.885
Lateral	82.2	89.4	74.9	0.915
Nasal	89.6	88.8	90.4	0.907
Open	84.5	82.0	86.9	0.841
Pause	93.3	96.0	90.5	0.958
Stop	85.0	83.3	86.7	0.855
Strident	92.7	92.2	93.2	0.932
Trill	85.1	98.0	72.2	0.986
Velar	86.2	91.7	80.6	0.930
Vocalic	84.2	82.1	86.3	0.841
Voice	88.0	86.7	89.2	0.885

This type of visualization could be useful for medical examiners to detect miss-pronunciation errors for the different groups of phonemes. The quality of the pronunciation for each group of phonemes in the phonological classes can be associated to the posterior probability in the posteriorgram. For instance, note the activation of the *nasal* at the beginning of the posteriorgram, which is related to the pronunciation of the phoneme /m/, or the activation of the *strident* at the end of the utterance, which is related to the pronunciation of the phoneme /s/.

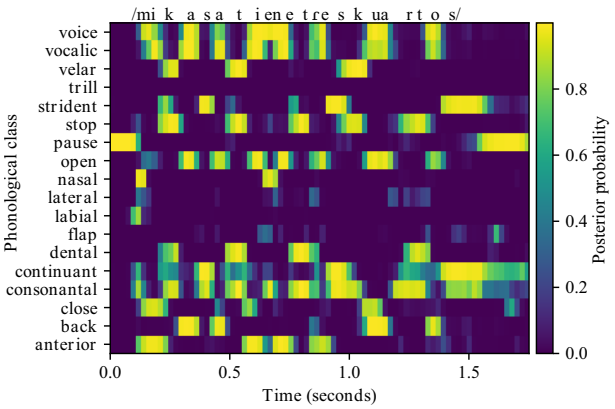

Figure 2: *Posteriorgram obtained for the Spanish sentence "mi casa tiene tres cuartos".*

A more detailed example is shown in Figure 3 for the vocalic, stop, nasal, and strident phonological posteriors in the Spanish sentence "mi casa tiene". Note how the nasal posterior is accurate to detect the phonemes /m/ and /n/, the strident posterior to detect the /s/ and the stop posterior to detect the /k/ and the /t/.

4.2. Phoneme recognition

The results of the proposed method to recognize the 21 phonemes of the Spanish language are shown in Table 5 in

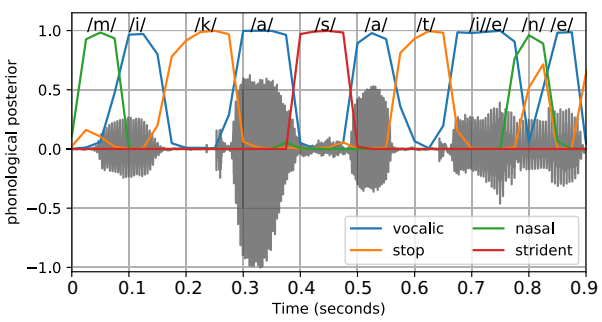

Figure 3: *Example of the vocalic, stop, nasal, and strident phonological posteriors estimated for the sentence "mi casa tiene" in Spanish language.*

terms of the precision, recall, and the F-score. The overall κ-index for this multi-class problem is 0.596. Note that there are phonemes classified with high precision such as the vowels /a/, /e/, and /o/, which could be explained due to the high percentage of occurrence of these three phonemes in the Spanish language [22]. The phonemes /s/ and /tʃ/ show higher recall than the others. This aspect gives insights about the capacity of the proposed model to learn and focus on the recognition of strident phonemes, which is reflected in the high accuracy obtained to recognize the *strident* phonological class (see Table 4).

Table 5: *Results for phoneme recognition using the proposed approach.* ***avg.*** *Average*

Phoneme	Precision	Recall	F-score
/a/	0.855	0.628	0.725
/e/	0.781	0.576	0.663
/i/	0.693	0.653	0.672
/o/	0.816	0.616	0.702
/u/	0.567	0.523	0.544
/b/	0.358	0.591	0.446
/d/	0.410	0.534	0.464
/f/	0.249	0.643	0.359
/g/	0.148	0.501	0.229
/x/	0.303	0.626	0.408
/k/	0.567	0.664	0.612
/l/	0.540	0.435	0.482
/m/	0.508	0.711	0.593
/n/	0.507	0.508	0.508
/p/	0.396	0.714	0.510
/ɾ/	0.371	0.369	0.370
/r/	0.240	0.656	0.351
/s/	0.705	0.812	0.754
/t/	0.557	0.739	0.635
/ʎ/	0.155	0.324	0.210
/tʃ/	0.503	0.873	0.639
/sil/	0.850	0.736	0.789
avg	0.674	0.625	0.637

The confusion matrix from Figure 4 shows with more detail the results for the phoneme recognition problem. Note that many of the miss-classification errors appear between similar phonemes, e.g., the 12% of the /u/ phonemes classified as /o/, the 16% of the /n/ phonemes classified as /m/, or the 11% of the /r/ phonemes classified as the phoneme /ɾ/. The most accurate phonemes to be recognized are the strident /s/ and /tʃ/, as it was explained previously. The most miss-classified phoneme is /ʎ/, which is confused mainly with /x/ and /i/.

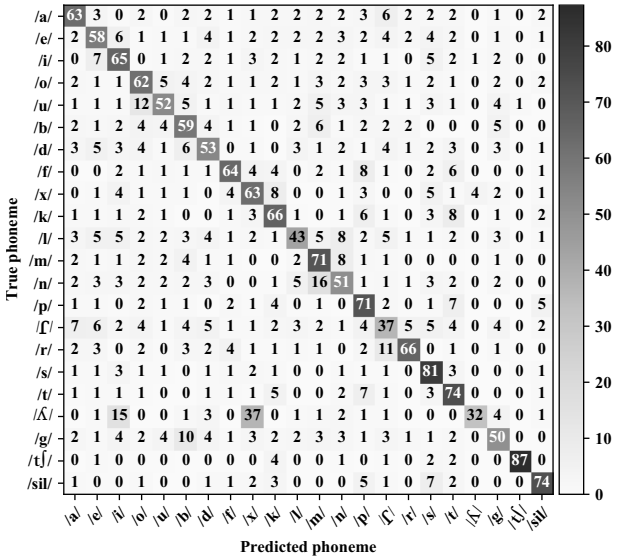

Figure 4: *Normalized confusion matrix (in %) for the phoneme recognition task using the proposed approach. The gray bar indicates the percentage of samples classified for each phoneme.*

5. Conclusion

We proposed a set of models to extract phonological posterior probabilities from speech. The models consists of a set of RNNs with bidirectional GRU units trained to recognize 18 phonological classes defined according to the mode and manner of articulation of the different phonemes of the Spanish language. The proposed models can be used to detect miss-pronunciation errors in different groups of phonemes, which can be highly useful in the assessment of pathological speech, or in the evaluation of non-native speakers. The accuracy of the proposed models ranges from 80.4% to 93.3%, depending on the phonological class. There are highly accurate models such as those to detect nasal, strident, or voice phonemes.

An additional model was considered to recognize individual phonemes of the Spanish language, in a multi-class scenario. The results suggest that there are phonemes such as /s/, /tʃ/, /m/, and /p/ that are detected accurately. On the other hand, many of the miss-classified phonemes are recognized as a similar phoneme according to the mode and manner of articulation.

The proposed models were trained to recognize phonological posteriors in Spanish language, but the training process can be easily adapted to other languages. Future models will include the estimation of phonological posteriors for the English and German languages. The trained models using Keras [13] are available as an open-source package that can be used to extract phonological posteriors directly from the speech signals[3].

6. Acknowledgments

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287. The authors also thank to CODI from University of Antioquia, grants # PRG2015-7683 and 2017-15530.

[3] https://github.com/jcvasquezc/phonet

7. References

[1] L. R. Rabiner, B. H. Juang, and J. C. Rutledge, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[2] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[3] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *Proceedings of ICASSP*, 2011, pp. 5884–5887.

[4] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of ICASSP*, 2013, pp. 6645–6649.

[5] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*, 2013.

[6] R. Gupta, T. Chaspari *et al.*, "Pathological speech processing: State-of-the-art, current challenges, and future directions," in *Proceedings of ICASSP*, 2016, pp. 6470–6474.

[7] Y. Jiao, V. Berisha, and J. Liss, "Interpretable phonological features for clinical applications," in *Proceedings of ICASSP*, 2017, pp. 5045–5049.

[8] M. Cernak *et al.*, "Characterisation of voice quality of Parkinson's disease using differential phonological posterior features," *Computer Speech & Language*, vol. 46, 2017.

[9] A. Asaei, M. Cernak, and M. Laganaro, "Paos markers: Trajectory analysis of selective phonological posteriors for assessment of progressive apraxia of speech," in *Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016, pp. 50–55.

[10] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *Proceedings of ICASSP*, 2015, pp. 992–996.

[11] M. Cernak and P. N. Garner, "Phonvoc: A phonetic and phonological vocoding toolkit," in *Proceedings of INTERSPEECH*, 2016, pp. 988–992.

[12] J. C. Vasquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease," *Journal of Communication Disorders*, vol. 76, pp. 21–36, 2018.

[13] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[14] J. L. Hieronymus, "Ascii phonetic symbols for the worlds languages: Worldbet," *Journal of the International Phonetic Association*, vol. 23, p. 72, 1993.

[15] A. R. Bagudanch, "Variation and phonological change: The case of yeísmo in Spanish," *Folia Linguistica*, vol. 51, no. 1, pp. 169–206, 2017.

[16] K. Irie, Z. Tüske, T. Alkhouli, R. Schlüter, and H. Ney, "LSTM, GRU, Highway and a Bit of Attention: An Empirical Overview for Language Modeling in Speech Recognition," in *Proceedings of INTERSPEECH*, 2016, pp. 3519–3523.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Deep Learning and Representation Learning Workshop*, 2014.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] C. D. Hernández-Mena and J. Herrera-Camacho, "Ciempiess: A new open-sourced mexican spanish radio corpus," in *Proceedings of the ninth international conference on language resources and evaluation (LREC14)*. European Language Resources Association (ELRA) Reykjavik, Iceland, 2014, pp. 371–375.

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

[21] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proceedings of the ICPhS*, 1999, pp. 607–610.

[22] M. Guirao and M. García-Jurado, "Frequency of occurence of phonemes in american Spanish," *Revue quebecoise de linguistique*, vol. 19, no. 2, pp. 135–149, 1990.