Interspeech 2018
2-6 September 2018, Hyderabad, India

Articulatory Feature Classification Using Convolutional Neural Networks

Danny Merkx and Odette Scharenborg

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
d.merkx@let.ru.nl, o.scharenborg@let.ru.nl

Abstract

The ultimate goal of our research is to improve an existing speech-based computational model of human speech recognition on the task of simulating the role of fine-grained phonetic information in human speech processing. As part of this work we are investigating articulatory feature classifiers that are able to create reliable and accurate transcriptions of the articulatory behaviour encoded in the acoustic speech signal. Articulatory feature (AF) modelling of speech has received a considerable amount of attention in automatic speech recognition research. Different approaches have been used to build AF classifiers, most notably multi-layer perceptrons. Recently, deep neural networks have been applied to the task of AF classification. This paper aims to improve AF classification by investigating two different approaches: 1) investigating the usefulness of a deep Convolutional neural network (CNN) for AF classification; 2) integrating the Mel filtering operation into the CNN architecture. The results showed a remarkable improvement in classification accuracy of the CNNs over state-of-the-art AF classification results for Dutch, most notably in the minority classes. Integrating the Mel filtering operation into the CNN architecture did not further improve classification performance.

Index Terms: Articulatory Features, Convolutional Deep Neural Networks, Mel Filtering

1. Introduction

The ultimate goal of our research is to improve an existing speech-based computational model of human speech recognition, Fine-Tracker, [1], on the task of simulating the role of fine-grained phonetic information in human speech processing. As part of this work we are investigating articulatory feature classifiers that are able to create reliable and accurate transcriptions of the articulatory behaviour encoded in the acoustic speech signal. Articulatory features (AFs), which are the acoustic correlates of articulatory events, have received a considerable amount of attention in automatic speech recognition (ASR) research [2-13] and are often considered as a solution to the problem of modelling the variation in speech using the standard 'beads-on-a-string' (i.e., using phones) paradigm [14]. Research has shown that the use of articulatory features (AFs), can help deal with the variability in the speech signal and improve the noise robustness of automatic speech recognition systems [2][5][6]. AFs have been used to improve computational models of human word recognition [1], build language independent phone recognisers [7] and, more recently, for multi-lingual ASR in low-resource settings [8]. Using multiple streams of information, that is, both acoustic features and articulatory features, has shown to improve speech recognition results for both HMM based and DNN based speech recognition systems [9][10].

Different approaches have been investigated for incorporating AFs into ASR systems, including support vector machines [4], hidden Markov models [2], and artificial neural, most notably multi-layer perceptrons, which have been successful at extracting AFs for many languages including Dutch, Czech, and German [1-7][11].

Recently, deep neural networks (DNNs) have been investigated for the task of AF classification, e.g., [8-13]. For instance, Badino and colleagues used DNN autoencoders to extract articulatory information, which combined with acoustic features showed a reduction in phone error rate in a DNN-HMM phone recogniser [12]. In [13], Siniscalchi et al. show that DNNs with just five layers led to remarkable improvements in AF classification over single hidden layer MLPs. Relatedly, deep convolutional neural networks (CNNs) have been successfully applied to the problem of phoneme recognition [15][16] and speech inversion [9]. In light of the success of deep CNNs, the first aim of this study is to investigate the application of deep CNNs for AF classification.

Secondly, in speech recognition, feature design and classifier design are often considered separate problems. For CNNs, the Mel Filterbank is a popular input feature (e.g., [16][17]). However, DNNs have been shown to be able to learn how to extract sensible features from data [18]. We propose a CNN architecture where the Mel filtering operation is included in the network as a convolutional layer, following Sainath et al. who showed an improvement in word recognition performance [19]. Training the acoustic feature extraction along with the rest of the network has the advantage that the features can be tuned to the classification task. This makes it possible to adjust the Mel Filterbanks to be optimal for each AF classifier individually. Furthermore this allows us to take a step backwards with regards to prepossessing and towards the raw speech signal leading to more plausible input for computational modelling tasks.

To summarise, this paper aims to improve AF classification by investigating two different approaches: 1. investigating the usefulness of a CNN for AF classification; 2. integration of the Mel filtering operation into the CNN architecture. In order to investigate the effect of these two approaches, the new models are compared with two sets of baseline models. The first set of baseline models are the original MLPs used in the Fine-Tracker computational model reported in [1]. Since DNNs are extremely data hungry, the amount of training material needed to train the CNNs is substantially larger than that used to train the MLPs in [1]. To be able to distinguish a possible effect of the DNN architecture from the possible effect of an increase in training material, a new set of MLPs with the same specifications as those in [1] was trained on the same amount of training material as the CNNs.

10.21437/Interspeech.2018-2275

Table 1: *Specification of the AFs, their types, and the number of hidden and output nodes in the MLPs and DNNs. The majority class for each AF is indicated in boldface.*

AFs	AF types	#hidden nodes	#output nodes
Manner	plosive, fricative, nasal, glide, liquid, **vowel**, retroflex, silence	300	8
Place	**nil**, bilabial, labiodental, alveolar, velar, glottal, palatal, silence	200	8
Voice	**+voice**, -voice	100	2
Backness	front, central, back, **nil**	200	4
Height	high, mid, low, **nil**	250	4
Rounding	+round, -round, **nil**	200	3
Dur-diphthong	long, short, diphthong, **silence**	200	4

2. Method

2.1. Materials

The speech material used in this study came from the Corpus Spoken Dutch (CGN, Corpus Gesproken Nederlands [20]). The material consisted of 64 hours of read speech by 324 unique speakers. The training data was split into a training (80% of the full data set), validation (10%) and test set (10%) with no overlap in speakers. AF labels were derived by first forced aligning the speech data with the phonemic transcriptions using a GMM-HMM system implemented in Kaldi [21]. Next, for each frame, the phonemic CGN label was replaced with the canonical AF types using a look-up table. The MLPs reported in [1] were trained on 3410 randomly selected utterances from the manually transcribed and segmented CGN read speech part (duration: 2h 50m). (Note that these sentences were also part of the data used to train and test the new baseline MLPs, however the manually checked phone segmentations were replaced by the forced alignments, in line with the rest of the speech material).

2.2. Articulatory features

We used the set of seven articulatory features from [1] as shown in Table 1. The names of the AFs are self-explanatory, except maybe for *dur-diphthong* which indicates whether a vowel is long, short, or is a diphthong (Dutch has three), and *backness* and *height* which indicate tongue position during the production of vowels. The majority class for each AF is indicated in boldface. *Nil* indicates 'not-specified', e.g., *place* of articulation is not specified for vowels.

2.3. Acoustic features

Three types of acoustic features were investigated. The two baseline MLPs were trained using 12 MFCCs and the log energy feature, augmented with the first and second derivatives resulting in 39 dimensional feature vectors. MFCCs were computed using 25 ms analysis windows with a 5 ms shift.

The CNN architecture was trained using Mel Filterbank features consisting of 64 filters, which were computed using 25 ms analysis windows with a 5 ms shift. The Mel Fbanks were computed according to the ETSI Distributed Speech Recognition Standard, with the only difference of using 64 filters instead of 23 [22].

For the extended CNN architecture, the Mel filtering operation is included in the network as a convolutional layer. The application of a Mel filter to the spectrogram is basically a one dimensional convolution over the entire frequency axis. The frequency spectral features are created by running the pipeline for the Mel Fbanks up to the fast Fourier transform (FFT) and stopping before the Mel filtering operation. The frequency spectral features were computed using 25 ms windows with 5 ms shift.

2.4. DNN architectures

The new baseline MLPs were implemented as described in [1]. The number of hidden nodes and output nodes for each AF are listed in Table 1. Each MLP had one hidden layer with a hyperbolic tangent non-linearity and a softmax output layer. The input to the MLPs consisted of the 39 dimensional MFCC features. Each training sample consisted of 11 consecutive frames with the middle frame being the frame to classify. The networks were trained using Nesterov momentum with a learning rate of 0.01 and momentum of 0.9, a learning rate decay of 0.5 per epoch and a batch size of 512. Training was stopped when the accuracy on the validation set started to drop. The baseline MLP of [1] is referred to as MLP-[1], while the baseline MLPs trained with the same amount of training material as the CNNs are referred to as Baseline MLPs.

The CNN architecture was implemented as described in [17]. The input features consisted of 11 consecutive frames of the Mel filtered signal for a total input size of 64 by 11. The architecture consisted of five blocks of two convolutional layers followed by a max pooling layer. The last block consisted of four fully connected layers with 2048 hidden nodes followed by a soft-max output layer. Figure 1 shows a visual representation of the architecture of the CNNs. The size of the convolutional filters was three by three throughout the network. The number of filters increased with depth, the first two convolutional layers had 64 filters, the next four layers had 128 filters and the last four layers had 256 filters. In the extended CNN architecture (referred to as CNN-Mf), the Mel filtering operation is implemented as a convolutional layer and inserted right after the input layer, see Figure 1. This layer consisted of 64 one dimensional filters with the weights preset to the filter coefficients of the Mel Fbanks. The input consisted of the 257 frequency spectral features for 11 consecutive frames.

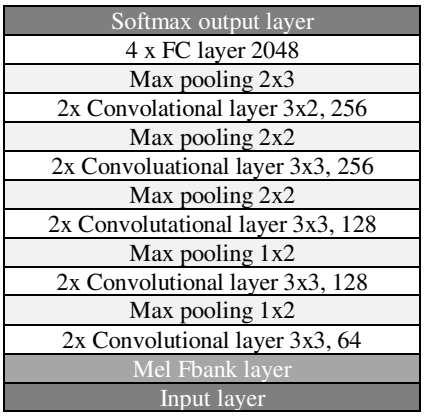

Figure 1: *CNN architecture. The Mel Fbank layer right after the input is absent in the case of the basic CNN and present in CNN-Mf.*

Table 2: *The classification results for the different neural network architectures for each AF. The best results for each AF are shown in boldface.*

	MLP-[1] baseline	MLP baseline	CNN	CNN-Mf	Chance
manner	76.6	74.0	**86.9**	84.6	29.8
place	76.2	72.7	**86.3**	84.9	29.81
voicing	89.3	90.2	**93.5**	92.9	53.5
backness	77.0	81.1	**89.2**	86.6	70.2
height	82.5	81.8	**89.2**	86.9	70.2
rounding	79.6	83.7	**90.6**	88.6	70.2
dur-diphthong	79.0	80.3	**88.2**	86.5	70.2

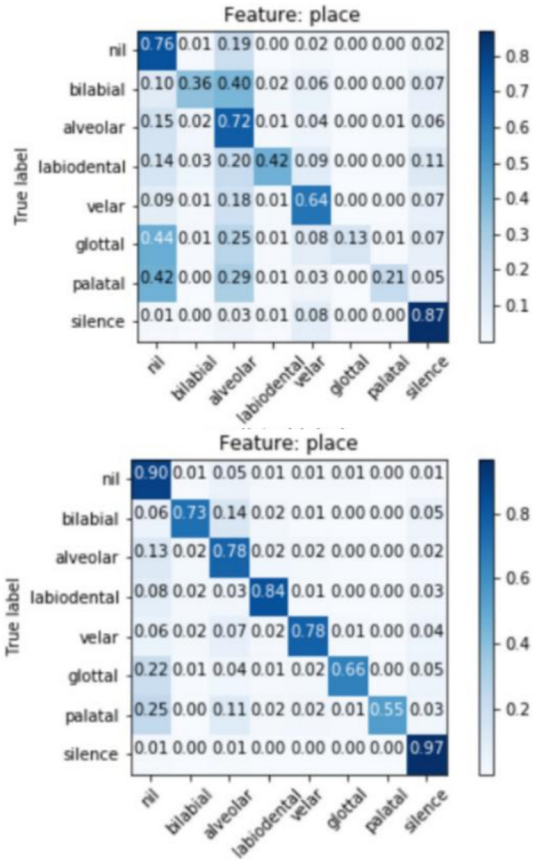

Figure 2: *Confusion matrices for the AF types of place of articulation of the MLP baseline model (top panel) and the CNN model (bottom panel).*

For both CNNs, spatial drop-out [23] was applied to the convolutions, randomly fixing 20% of the output feature maps to 0 during training. The fully connected layers received a regular drop-out of 40% of the nodes. All layers have a ReLU non-linearity and batch normalisation applied before the non-linearity [23]. The CNN networks were trained using Nesterov momentum with a learning rate of 0.01 and momentum of 0.9, a learning rate decay of 0.5 per epoch and a batch size of 512. The networks were trained for five epochs.

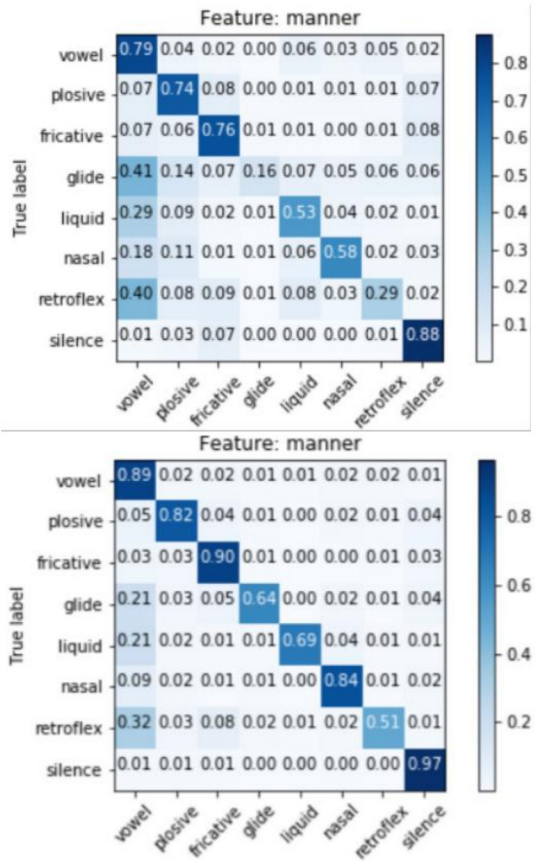

Figure 3: *Confusion matrices for the AF types of place of articulation of the MLP baseline model (top panel) and the CNN model (bottom panel).*

3. Results

Table 2 presents the classification results for the two baseline systems and the new DNN architectures. Since the number of training frames was not equally distributed over the different AF types, we also report the chance level performance for each AF, which equals the relative size of the majority class for each AF type (see Table 1). Each AF except *voicing* had a clear majority class. As shown in Table 2, all systems performed well above chance level on all AFs.

The new MLP baseline trained with the increased amount of training material did not substantially outperform the MLPs reported in [1]. The new MLP baseline performed better on *backness* and *rounding*, but worse on *manner* and *place of articulation* with minor differences on the other AFs. So, simply using more training data did not improve AF classification.

The basic CNN architecture showed a large improvement over the MLP baselines, with relative improvements over the MLP baseline of up to 18.61%. These improvements were thus due to the change in architecture of the neural nets and not due to the increase in amount of training material. Integrating the Mel filtering operation into the CNN architecture, however, did not further improve AF classification results. The performance of the extended CNN-Mf architecture was slightly worse than that of the basic CNN architecture with deteriorations ranging between 0.6% and 2.6% absolute.

Large class imbalances may cause the classifiers to become biased towards the majority class, both during training

and testing. We therefore also investigated the classification performance of the AF types separately. The bias towards the majority class was largest for the *place* feature (i.e., *nil*) and the *manner* feature (see also the chance levels in Table 2; i.e., *vowel*). Figures 2 and 3 show the confusion matrices for *place* and *manner of articulation*, respectively, with the top panel showing the new Baseline MLP results and the bottom panel showing the results of the best performing CNN architecture. Darker shades indicate higher classification rates; while the diagonal indicates correct classification for each AF type.

The MLP baselines showed the worst bias to the majority class, which shows up as a vertical column in the Figures. For *place*, four of the eight AF types were more often misclassified than correctly classified: for *glottal* and *palatal*, the confusion with the majority class *nil* was higher than their accuracy. While there are still some confusions with the majority class for the CNN architectures, the confusion matrices show that all classes are more often correctly classified than misclassified, with especially the *glottal* and *palatal* classes showing clear increases in accuracy. Similar results were found with both CNN architectures; a decreased bias towards the majority class and an increase in classification accuracy compared to the MLP baseline for every single AF type.

4. Discussion

The aim of this study was two-fold: improve articulatory feature classification by using deep CNNs, and investigate the integration of the Mel filtering operation into the CNN architecture, in order to obtain improved AF representations for the computational model Fine-Tracker [1]. To this end, three types of neural networks were trained. First, a new MLP baseline was trained according to the architecture of the original MLPs in the Fine-Tracker model [1] in order to account for the necessary increase in amount of training material to train the deep CNNs. Although using more training data typically increases the performance and generalisability of DNNs, somewhat unexpectantly, the results showed that simply using more training data did not substantially increase the AF classification performance of the new MLP baseline compared to the MLPs reported in [1]. A possible explanation is that the number of hidden nodes for each AF was optimised for the dataset used in [1] through tuning experiments. The optimal number of hidden nodes may be different for the dataset used in the current study. Importantly, these results show that an increase in performance of the CNNs over the MLPs is due to a change in the architecture of the models.

Secondly, a deep CNN was implemented following the architecture of [17]. Thirdly, an extension to this network was implemented in which part of the pre-processing was integrated as a convolutional layer. Both CNN architectures were a clear improvement over the baseline MLPs. The basic CNN gave the best classification results for all AFs with relative improvements of up to 18.61% over the baseline. While the basic CNN outperformed the extended CNN-Mf, the differences were only minor. These differences could be due to differences in the initialisation of the network weights. Follow-up research will have to answer this question. We noticed that the basic CNN architecture was almost at the top performance after even the first epoch, further training led to only small improvements. The performance of the CNN-Mf architecture was relatively low after the first epoch but increased steadily with further training. Increasing the depth of the network might simply require a longer training time. Our results are competitive with other neural network approaches. For instance in [26], frame accuracies of 85.0% and 72.5% are reported for *manner* and *place* of articulation respectively using recurrent neural networks (RNNs).

Six of our seven AFs had clear majority classes. An investigation of the AF classification confusion matrices showed that the classifiers were biased towards their majority class. This bias was so large for the MLPs that some classes were more often misclassified as the majority class than they were correctly classified. Interestingly, the best performing CNNs showed larger improvements in classification accuracy for the minority classes (i.e., the AF types with lower frequency of occurrence in the data) than for the majority classes. The bias towards the majority class was substantially decreased for the CNNs where all classes were more often classified correctly than misclassified, and where confusion with the majority class decreased for all AF types. This is an important improvement in the quality of the AF classifiers. For instance, the majority class for *backness*, *height*, *rounding* and *dur-diphthong* is *nil*, a label assigned to all consonants and silence. While distinguishing between vowels on the one hand and consonants and silence on the other is necessary in order to be able to distinguish between the different characteristics of vowels, the interesting information such as vowel *height* is actually captured by the minority labels. If these classifiers are biased towards the majority class they are all fairly accurate at detecting consonants but they provide less accurate information about the characteristics of the vowels. In this sense the CNNs are even more of an improvement than the overall accuracy would suggest.

Recent work using Bi-directional LSTMs has shown the potential to reach even higher accuracies than those reported here, see e.g., [26]. The bi-directionality of these networks, however, excludes AFs derived in this way as plausible features for computational modelling as in the backwards direction these networks are allowed to use information from the future, which is not in line with how human listeners process speech. However, we think that CNNs topped with unidirectional LSTMs are a promising direction for future research, combining the CNNs ability to capture information in the frequency spectrum and the LSTMs ability to capture temporal dependencies.

5. Conclusion

We presented the, to the best of our knowledge, first application of CNNs to the task of AF classification. The results showed a remarkable increase in AF classification accuracy compared to the state-of-the-art for Dutch. The improvements are most notably found in an increase in accuracy for the minority classes and a reduction of the classification bias to the majority classes. Integrating the Mel filtering operation into the CNN architecture did not further improve classification performance. The next step is to integrate the improved AF classifications in the computational model of human spoken-word recognition in order to investigate the effect of improved AF classification on the computational model's simulation power.

6. Acknowledgements

This work was carried out by D. Merkx as part of an MSc project under the supervision of O. Scharenborg. O. Scharenborg was supported by a Vidi-grant from NWO (grant number: 276-89-003).

7. References

[1] O. Scharenborg, "Modelling the use of durational information in human spoken-word recognition," *Journal of the Acoustical Society of America,* vol. 127, no. 6, pp. 3758–3770, 2010

[2] K. Kirchhoff, "Robust speech recognition using articulatory information," *Ph.D. thesis, University of Bielefield*, 1999.

[3] M. Wester, "Syllable classification using articulatory-acoustic features," in *Eurospeech 2003 – 8th European Conference on Speech Communication and Technology, September 1-4, Geneva, Switserland, Proceedings,* 2003, pp. 233–236.

[4] O. Scharenborg, V. Wan and R.K. Moore. "Capturing fine-phonetic variation in speech through automatic classification of articulatory features", in *SRIV 2006 – International Workshop on Speech Recognition and Intrinsic Variation, May, Toulouse, France, Proceedings,* 2006 pp. 77–82.

[5] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, pp. 723–42, 2007.

[6] K.. Kirchhoff, G. A. Fink and G. Sagerer, „Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319, 2002.

[7] S. M. Siniscalchi, T. Svendsen and C.-H. Lee, "Towards a detector-based universal phone recognizer," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 31 - April 04, Las Vegas, U.S.A.*, *Proceedings,* 2008, pp. 4261–4264.

[8] M. Müller, S. Stüker and A. Waibel, "Towards Improving Low-Resource Speech Recognition Using Articulatory and Language Features", in *IWSLT 2016 – 13th International Workshop on Spoken Language Translation (IWSLT), December 8-9, Seattle, USA, Proceedings,* 2016, pp

[9] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman and M. Tiede, "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Communication*, vol. 89, pp. 103–112, 2017.

[10] V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Richey, J. Yuan and M. Liberman , „Articulatory trajectories for large-vocabulary speech recognition," *Acoustics, Speech, and Signal Processing*, pp 7145–7149, 2013.

[11] P. Mizera and P. Pollak, "Robust neural network-based estimation of articulatory features for Czech," *Neural Network World*, vol. 24, pp. 463–478, 2014

[12] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "Integrating Articulatory Data in Deep Neural Network-based Acoustic Modeling," *Computer Speech & Language*, vol. 36, 2015.

[13] S. M. Siniscalchi, D. Yu, L. Deng and C. Lee, "Exploiting deep neural networks for detection-based speech recognition," *Neurocomputing*, vol. 106, pp. 148–157, 2013.

[14] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *1999 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 12-15, Keystone, CO, Proceedings,* 1999, pp. 79–84.

[15] C. Shulby, M. Dais Ferreira, R. Mello and S. Aluisio, "Acoustic Modeling Using a Shallow CNN-HTSVM Architecture," arXiv:1706.09055v1 [cs.SD] 27 Jun 2017.

[16] O. Abdel-Hamid, A. Mohamed, H. Jiang and G. Penn, "Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition," in *2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 25-30, Kyoto, Japan, Proceedings*, 2012, pp 4277–4280.

[17] Y. Qian and P. Woodland, "Very deep convolutional neural networks for robust speech recognition", arXiv:1610.00277v1 [cs.CL] 2 Oct 2016.

[18] Y. LeCun, "Learning Invariant Feature Hierarchies", in *Fusiello A., Murino V., Cucchiara R. (eds) Computer Vision – ECCV 2012. Workshops and Demonstrations. ECCV 2012. Lecture Notes in Computer Science, vol 7583. Springer, Berlin, Heidelber.* 2012

[19] T. Sainath, B. Kingsbury, A. Mohamed and B. Ramabhadran , "Learning filter banks within a deep neural network framework," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, (ASRU), December 8-12, Olomouc, Czech Republic, Proceedings*, 2013, pp. 297–302.

[20] N. H. J. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.-P. Martens, M. Moortgat and H. Baayen, "Experiences from the Spoken Dutch Corpus project," in *LREC – Third International Conference on Language Resources and Evaluation, May 29-31, Las Palmas de Gran Canaria*, *Proceedings,* 2002, pp. 340–347.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer and K. Vesel, "The Kaldi speech recognition toolkit", in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 11-15, Waikoloa, U.S.A., Proceedings*, 2011.

[22] European Telecommunications Standards Institute, "ETSI ES 201 108 V1.1.3." , 2003, Retrieved on 20/03/2018. [Online]. Available: www.etsi.org

[23] J. Tompson, R. Goroshin, A. Jain, Y. Lecun and C. Bregler, "Efficient object localization using Convolutional Networks," arXiv:1411.4280v3 [cs.CV] 9 Jun 2015.

[24] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37

[25] K. Hacioglu, B. Pellom and W. Ward, "Parsing speech into articulatory events," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 17-21, Montreal, Canada, Proceedings*, 2004, pp. I-925-8, vol. 1.

[26] M. Müller, S. Stiiker and A. Waibel, "DBLSTM based multilingual articulatory feature extraction for language documentation," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, *December 16-20, Okinawa, Japan, Proceedings,* 2017, pp. 417–423.