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

Using Speech Production Knowledge for Raw Waveform Modelling based Styrian Dialect Identification

S. Pavankumar Dubagunta[1,2], *Mathew Magimai Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland

Abstract

This paper addresses the Styrian Dialect sub-challenge of the INTERSPEECH 2019 Computational Paralinguistics Challenge. We treat this challenge as dialect identification with no linguistic resources/knowledge and with limited acoustic resources, and develop end-to-end raw waveform modelling based methods that incorporate knowledge related to speech production. In this direction, we investigate two methods: (a) modelling the signals after source system decomposition and (b) transferring knowledge from articulatory feature models trained on English language. Our investigations show that the proposed approaches on the ComParE 2019 Styrian dialect data yield systems that perform better than low level descriptor-based and bag-of-audio-word representation based approaches and comparable to sequence-to-sequence auto-encoder based approach.

Index Terms: Source-filter decomposition, articulatory modelling, raw-waveform modelling, convolutional neural networks, computational paralinguistics.

1. Introduction

Dialect identification (DID) involves distinguishing the acoustic, pronunciation and grammatical variations within a language used by people usually from different demographic regions. It is useful in customising automatic speech recognition systems which underperform due to changes in the dialects, in identifying a person's regional origin and ethnicity in forensic analysis [1], in tailoring speech synthesis systems for improved user experience and so on.

DID is generally approached from the linguistic differences. DID is considered harder to solve than language identification, as dialectal differences within a language are generally more subtle than those between languages. To cite a few works in this direction, [2] proposed to learn phonetic rules for DID, [3] learned n-gram statistics of phones and used lattice rescoring for DID, [4] performed phonotactic based DID using convolutional neural networks (CNNs).

In terms of the acoustic differences, DID is closely related to accent identification from speech. In this direction, some works used i-Vectors and bottleneck features [5] from acoustic data, Eigen channel modelling based on factor analysis [6], Gaussian mixture model (GMM) based supervectors representing phone segments [7, 1]. Traditional feature sets for DID included shifted delta coefficients [8].

DID without linguistic resources and using limited speech data has been less studied. In [8], the authors addressed the lack of linguistic resources by using unsupervised learning, i.e. by first modelling data using GMMs and thereby training neural networks to predict the posterior probabilities of these unsupervised models and through further processing. Other works such as [9, 10] also focused on unsupervised representations for DID. In [11], an end-to-end framework from features such as spectrogram was proposed in an unsupervised learning setting, based on factorised hierarchical variational autoencoders. They address the acoustic resource scarcity by augmenting data through speed perturbation [12], which varies the speed of the original signals to create two more of their variants at 0.9 and 1.1 relative speeds.

This paper is towards the ComParE 2019 challenge [10], where each short utterance is to be classified into one of the three Styrian dialects: the Northern, Eastern and Urban variants. We approach this problem as DID with limited acoustic resources and with no linguistic resources. Since these dialects lack distinction in their pitch accent, we hypothesise that most of the distinguishable information must be in the differences related to the shape of the vocal tract during articulation. In this direction, we study DID through CNN-based modelling of signals of raw speech that contain vocal-tract related information and through modelling their articulatory differences. Specifically, (a) motivated by the recent works that modelled raw speech directly for specific tasks [13, 14, 15, 16] and their signal-processed variants [17], we study DID using knowledge from source-filter decomposition, and (b) motivated by the acoustic modelling of articulatory differences among phones [18], we study learning such knowledge from raw signals of speech on resource-rich English language and transferring it to the Styrian DID problem.

The rest of the paper is organised as follows: §2 summarises the data set information and the baselines set by the challenge, §3 describes the source-filter decomposition based approach, its experimental setup and results, §4 describes the articulatory based approach and its experimental results, §5 discusses the further analysis and findings and finally §6 concludes the paper.

2. Data set and baselines

The ComParE 2019 Styrian dialect data set [10] consists of 76 minutes of training data, with short utterances of average duration 0.87 seconds, 34 minutes of development[1] data and 29 minutes of test data. Except for the dialect labels of the training and dev sets, no other linguistic resources (such as transcriptions) were available. The labels of the test set were reserved with the organisers, and only 5 systems from each participating team were allowed to be submitted for the test set score evaluation.

The baseline systems provided include support vector machine (SVM) based classifiers trained on: (a) multiple features related to the voice source and the vocal tract parameters

e-mail: (see http://www.idiap.ch/en/people/directory).

[1] Development set is referred to as the *dev* set hereafter.

http://dx.doi.org/10.21437/Interspeech.2019-2398

and their statistical properties, known as low-level descriptors (LLD), (b) histogram representations of clustered LLDs, known as bag-of-audio-words (BoAW), (c) feature representations from sequence-to-sequence auto-encoders (S2SAE) trained on Mel-spectrograms. Table 2.1 summarises their best performances, obtained by parameter tuning, in terms of unweighted average recall (UAR) % on all the classes. The last column uses a fusion of S2SAE models trained using different signal-to-noise ratio thresholds; it is worth noting that the systems are sensitive to parameter variations, as observed in the test set scores.

Table 2.1: *Baseline systems provided with the challenge and their UAR% on the dev and test sets.*

	Best system			Fused
Data set	LLD	BoAW	S2SAE	S2SAE
dev	38.8	38.2	46.7	45.9
test	36.0	32.4	47.0	35.5

3. Source-filter decomposition based DID

We adopt the CNN-based framework described in [13], which was initially developed for speech recognition and later extended to other tasks [14, 19]. As illustrated in Fig. 3.1, the proposed system takes as input a fixed length signal and processes it through multiple convolutional layers followed by fully connected layers and outputs the probabilities of observing each of the three dialects. During testing, such posterior probability vectors are averaged across each utterance and a decision is made at the utterance-level based on the highest probability. Our recent work [17] showed that filtering raw speech based on prior knowledge facilitates better task-specific modelling. Along similar lines, apart from directly modelling raw speech and using speed perturbation (SP) as suggested in [11] to partially address data scarcity, we propose to use the following signal processing techniques to extract signals rich in vocal-tract related information: (a) homomorphic filtering and (b) linear prediction based filtering.

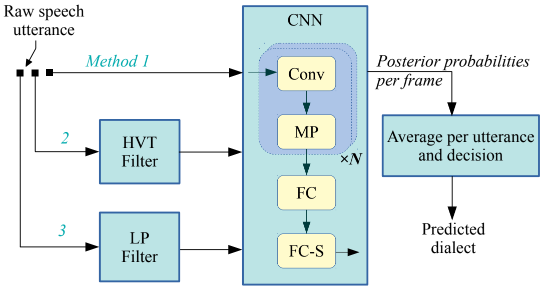

Figure 3.1: *Proposed approach based on CNNs and using source-filter decomposition. Conv: convolutional layer with rectified linear (ReLU) activation, MP: max-pooling, FC: fully connected layer with ReLU activation, FC-S: FC layer with softmax activation.*

3.1. Homomorphically filtered vocal-tract filter (HFVT) signal

Complex cepstrum transforms convolutive components of a time-domain signal into additive components. Here we employ a simple low-pass cepstral lifter to approximately remove the fast varying voice source component from a speech signal and to retain the vocal tract response [20]. Since the complex cepstrum transform is invertible, a corresponding time domain signal can be constructed from the liftered cepstrum. We perform this analysis using a sliding window on each raw speech utterance, and by overlap-adding the resultant vocal tract signals to get the corresponding HFVF signal of each utterance.

3.2. Linear prediction estimated (LPE) signal

Linear prediction (LP) technique models the predictable components in a signal [21]. For a typical short quasi-stationary segment of speech, this corresponds to fitting the formant-related vocal-tract structure. However, the quality of such modelling depends on the order of the LP used. The difference signal between the original and the estimated signal, more popularly known as the LP residual, contains more of the unpredictable vocal-source related information. In various applications such as telephony, the LP residual signals are constructed per utterance by overlap-addition over short segments of speech. In a similar manner, the LP estimated signals can be overlap-added to construct utterance-level LPE signals.

3.3. Experimental setup

HFVT and LPE signals were generated using MATLAB. HFVT signals were extracted with a 40ms Hanning window, shifted by 20ms and were liftered at 50 sample quefrency cut-off. LPE signals were predicted using 30ms Hamming windows, shifted by 10ms and using 12^{th} order LP modelling. SP was performed using SoX tool at additional 0.9 and 1.1 speeds. As discussed earlier, these signals were computed at utterance level through overlap and add, and were then processed through CNNs.

CNN systems were trained using Keras deep learning library [22] with Tensorflow backend [23]. The architecture used is listed as *SigProc* in the Table 3.1. In order to avoid skewed results and to make the systems more robust to various random initialisations, 5-fold cross-validation was conducted using leave-one-out method. In other words, 5 CNN systems were trained for each experiment by cross-validating on a left-out unseen part of the training set. During training, all the three classes were ensured of equal representation in each epoch by duplicating some of the utterances presented. “FC” layer in this architecture contains 100 nodes. The input to the CNNs is a 250ms signal, overlapped by a 10ms shift; these parameters are inspired from earlier works such as [13, 17]. The targets to the CNNs are one-hot encodings of the dialects. The networks were trained using cross-entropy loss with stochastic gradient descent. Learning rate was halved, in the range 10^{-2} to 10^{-6}, between successive epochs whenever the validation-loss stopped reducing. The posterior probabilities obtained from the 5 CNNs for each utterance were averaged before classification.

3.4. Results

Table 3.2 summarises the results of the proposed methods. It is worth noting that the HFVT and LPE methods without SP perform better than raw speech with SP on the dev set, and perform comparable on the test set. §5.3 gives a detailed analysis on SP.

Table 3.1: *CNN architectures. N_f: number of filters, kW: kernel width, dW: kernel shift, MP: max-pooling.*

Model (Input frame size)	Layer	Conv			MP
		N_f	kW	dW	
SigProc (250ms)	1	128	30	10	2
	2	256	10	5	3
Artic (250ms)	1	80	30	10	3
	2,3	60	7	1	3

Table 3.2: *UAR% on the dev and test sets. RawSP indicates raw speech using speed perturbation.*

Data set	RawSP	HFVT	LPE
dev	44.2	**46.8**	46.3
test	**34.2**	34.0	**34.2**

4. Transfer learning linguistic knowledge

Since the Styrian DID challenge data lack linguistic resources, in this section we describe our approach to incorporate such knowledge from another resource-rich language into the current task. The typical linguistic sub-word units of any language, such as phonemes, can be mapped to the articulatory properties of the vocal apparatus that cause to produce the associated sounds. Such properties include the place of constriction, the height of the tongue, roundedness of the lips and so on. When such a knowledge exists in a language, mappings from the linguistic units to their articulatory feature (AF) representations can be learned [18]. Such AF representations were shown to be useful for improved pronunciation modelling, noise robustness and multi-lingual portability. Along these lines, our contribution is two fold: (a) motivated by better task-specific modelling through automatic feature learning, we propose to model AFs directly from raw speech and its variants HFVT and LPE, using phone alignments and phone-to-AF mappings, and (b) as shown in Fig. 4.1, we propose to utilise such AF model parameters in the Syrian DID task, primarily as model initialisers and further as feature embeddings.

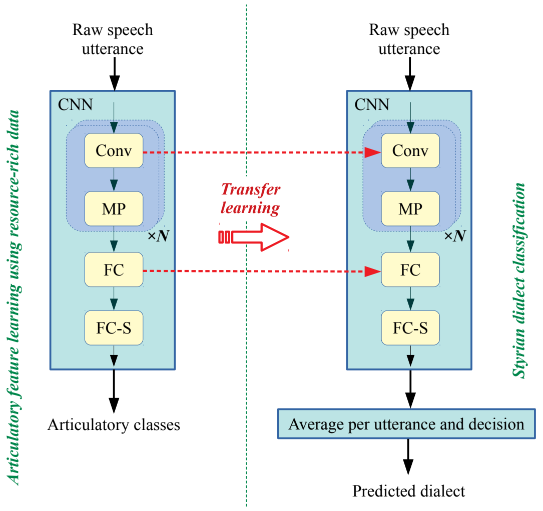

Figure 4.1: *Transfer learning based on articulatory knowledge.*

4.1. Experimental setup

For training the articulatory networks, AMI [24] corpus recorded using independent headset microphone was used, which consists of 77 hours of meetings. Kaldi setup was used to train hidden Markov models (HMMs) for context-dependent phones, where the HMM states were jointly modelled by using subspace GMMs. The corresponding frame-to-phone alignments were used to train the AF CNNs, using phone-to-AF mappings from Table B.1 of [18]. More specifically, for each kind of the input signal, 4 AF CNNs were trained to individually predict the 4 AF categories: place, manner, height and vowel. The model architecture used, listed as *Artic* in the Table 3.1, was inspired from raw-speech based phone classification first proposed in [13]. "FC" layer in this architecture contains 1024 nodes. AF training was performed on a 70 hour *clean* subset of the training set, which is a standard practice followed in the Kaldi recipe.

Transfer learning to Styrian DID involved using the 4 AF-CNNs to initialise another 4 corresponding CNNs, for DID, of the same architecture (*Artic*) except for the output classification layer. For transfer learning as feature embeddings, refer to §5.1. Training was performed using the training set, by using a decaying learning schedule as described in §3.3, and by cross-validating on the entire training set at the end of each epoch. The posterior probabilities obtained from the 4 CNNs for each utterance were averaged before classification.

4.2. Results

Table 4.1 summarise the results using initialisation based transfer learning. It can be observed that raw speech with SP performs better on the test set than the other proposed approaches and the existing LLD and BoAW based baselines. In comparison with the S2SAE baseline, the method performs better than the fused S2SAE baseline. Also, the LPE signal based systems (without SP) perform better on the dev set than all the existing baselines and the proposed appproaches. We remind the reader that the submissions on the test set were capped to 5, and hence we could not evaluate the HFVT *Artic* system on the test set. Finally, on comparing Tables 3.2 and 4.1, it can be observed that the proposed *Artic* methods perform generally better than the *SigProc* methods, especially on the test set. This validates that AF based transfer learning through initialisation helps in improving the Styrian DID.

Table 4.1: *Effect of initialisation based transfer learning from AF-CNNs on the UAR%.*

Expt.	dev					test
Artic	Place	Manner	Height	Vowel	Fused	Fused
RawSP	43.5	44.4	45.3	43.0	46.6	**36.6**
HFVT	44.0	44.2	45.0	44.3	45.0	-
LPE	45.5	47.3	46.2	45.0	**47.0**	35.6

5. Analysis

In this section, we analyse the following aspects. First, we study transfer learning using AF embeddings and how such methods can be useful to assess the most distinctive characteristics of articulation among the dialects. Second, we study whether the voice-source plays a role in Styrian DID, by using the proposed source-filter decomposition based methods to extract voice-source related components. Finally, we study how SP affects the proposed approaches.

5.1. Transfer learning using AF embeddings

The study summarised in Table 4.1 is limited to parameter initialisation based transfer learning. However, in such approach, the parameters of all the layers except for the final layer can be freezed during the training process that classifies Styrian dialects. This corresponds to extracting articulatory embeddings trained on the resource-rich language and using them to build a linear classifier for the Styrian DID problem. Table 5.1 shows such results, where the experimental setup is identical to that of §4.1 except for the parameter freezing. Although the results suggest that updating all the parameters gives an improved classification, they indicate that *manner* of articulation may carry the most distinguishable information among the other AFs for Styrian DID.

Table 5.1: *Effect of AF-embedding based transfer learning on the UAR% of the dev set.*

Expt. Artic	Place	Manner	Height	Vowel	Fused
RawSP	40.0	43.8	42.3	42.8	42.4
HFVT	43.5	43.1	44.6	44.1	44.7
LPE	39.3	45.8	42.2	41.8	43.8

5.2. The contribution of voice-source

Closely following the approaches in §3, we extracted the two signals: (a) homomorphically filtered voice source (HFVS), by negating the cepstral lifter described in §3.1, and (b) by overlap-adding the LP residual (LPR) as described in §3.2. Table 5.2 shows the results of modelling these signals using the *SigProc* CNN architecture. It can be clearly seen that the vocal-tract related differences distinguish Styrian dialects better than those of the voice-source.

Table 5.2: *Effect of the voice-source component on the UAR% of the dev set.*

Data set	HFVS	LPR
dev	35.2	37.1

5.3. The effect of speed perturbation

Table 5.3 shows several experiments with and without SP. It can be seen that raw speech based methods benefit from SP, as also indicated in [11], whereas the other proposed signals indicate no such gains.

Table 5.3: *Effect of SP on the UAR% of the dev set.*

Method		Raw	HFVT	LPE
Source-filter	-	41.8	46.8	46.3
	SP	44.2	45.6	44.9
AF	-	45.0	45.0	47.0
	SP	46.6	45.6	46.2

6. Conclusions and future directions

In this paper, we proposed raw waveform modelling approaches based on source-filter decomposition, viz. by HFVT and LPE signal extraction, and using AF based transfer learning for Styrian dialect identification. These approaches were shown to perform better than the existing LLD and BoAW based methods and comparable to the S2SAE based approach. Furthermore, the AF based transfer learning approach was shown to achieve better modelling when linguistic resources are unavailable. The paper also showed that the vocal-tract related differences play a better role in distinguishing Styrian dialects than the voice-source, particularly in terms of the manner of articulation. Finally, the source-filter decomposition methods were shown to yield competent systems without data augmentation.

In the future, we plan to investigate how the proposed methods such as source-filter decomposition and AF based transfer learning can infer assessment in other tasks, for instance in natural versus synthetic speech classification. We would also like to investigate the proposed approaches on DID in languages with more acoustic resources. Finally, we plan to systematically study the efficacy of modelling AFs from raw speech for tasks such as speech recognition, as compared to the existing feature extraction based methods.

7. Acknowledgements

This work was funded by the Hasler Foundation under the project Flexible Linguistically-guided Objective Speech aSsessment (FLOSS). The authors gratefully thank them for their financial support.

8. References

[1] F. Biadsy, J. Hirschberg, and M. Collins, "Dialect recognition using a phone-gmm-supervector-based svm kernel," in *Proceedings of Interspeech*, 2010, pp. 753–756.

[2] N. F. Chen, W. Shen, J. P. Campbell, and P. A. Torres-Carrasquillo, "Informative dialect recognition using context-dependent pronunciation modeling," in *Proceedings of ICASSP*, May 2011, pp. 4396–4399.

[3] R. Tong, B. Ma, H. Li, and E. S. Chng, "Target-aware lattice rescoring for dialect recognition," in *Proceedings of Interspeech*, 2011, pp. 733–736.

[4] M. Najafian, S. Khurana, S. Shan, A. Ali, and J. Glass, "Exploiting convolutional neural networks for phonotactic based dialect identification," in *Proceedings of ICASSP*, 2018, pp. 5174–5178.

[5] A. Ali, N. Dehak, P. Cardinal, S. Khurana, S. H. Yella, J. Glass, P. Bell, and S. Renals, "Automatic dialect detection in arabic broadcast speech," in *Proceedings of Interspeech*, 2016, pp. 2934–2938. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1297

[6] Y. Lei and J. H. Hansen, "Factor analysis-based information integration for arabic dialect identification," in *Proceedings of ICASSP*, 2009, pp. 4337–4340.

[7] F. Biadsy, J. Hirschberg, and D. P. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Proceedings of Interspeech*, 2011, pp. 745–748.

[8] Q. Zhang and J. H. Hansen, "Dialect recognition based on unsupervised bottleneck features." in *Proceedings of Interspeech*, 2017, pp. 2576–2580.

[9] Q. Zhang and J. H. L. Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 5, pp. 873–882, May 2018.

[10] B. W. Schuller *et al.*, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," submitted to Interspeech, http://www.compare.openaudio.eu/wp-content/uploads/2019/03/INTERSPEECH_2019_ComParE.pdf, 2019, [Online; accessed 22$^{\text{nd}}$ March 2019].

[11] S. Shon, A. Ali, and J. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," *arXiv preprint arXiv:1803.04567*, 2018.

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of Interspeech*, 2015, pp. 3586–3589.

[13] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proceedings of Interspeech*, 2013, pp. 1766–1770.

[14] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proceedings of ICASSP*, 2018, pp. 4884–4888.

[15] ——, "On learning vocal tract system related speaker discriminative information from raw signal using CNNs," in *Proceedings of Interspeech*, 2018, pp. 1116–1120.

[16] S. H. Kabil, H. Muckenhirn, and M. Magimai.-Doss, "On learning to identify genders from raw speech signal using CNNs," in *Proceedings of Interspeech*, 2018, pp. 287–291.

[17] S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss, "Learning voice source related information for depression detection," in *Proceedings of ICASSP (accepted)*, 2019. [Online]. Available: http://publications.idiap.ch/downloads/papers/2019/Dubagunta_ICASSP-2_2019.pdf

[18] R. Rasipuram and M. Magimai.-Doss, "Articulatory feature based continuous speech recognition using probabilistic lexical modeling," *Computer Speech and Language*, vol. 36, pp. 233–259, 2016. [Online]. Available: http://publications.idiap.ch/downloads/papers/2015/Rasipuram_CSL_2015.pdf

[19] B. Vlasenko, J. Sebastian, D. S. Pavan Kumar, and M. Magimai.-Doss, "Implementing fusion techniques for the classification of paralinguistic information," in *Proceedings of Interspeech*, 2018, pp. 526–530.

[20] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*. Pearson, 2011.

[21] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

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

[23] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," http://tensorflow.org/, 2015.

[24] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proceedings of International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.