IMPROVING PHONEME SET DISCOVERY FOR DOCUMENTING UNWRITTEN LANGUAGES *

Markus Müller, Jörg Franke, Sebastian Stüker, Alex Waibel
Karlsruhe Institute of Technology
m.mueller@kit.edu

Abstract: Many of the 7,000 living languages in the world are currently threatened by extinction. In order to preserve these languages and the cultural heritage linked with them, they need to be documented. This is a challenging and time consuming task, even for trained specialists. Helping linguists in language documentation is the goal of the French-German ANR-DFG project BULB. The first step in documenting a language is the discovery of the phonetic inventory. We aim at assisting linguists during this step by proposing a segmentation of audio data into phoneme-like units and by clustering these units using articulatory features. In this work, we refine our existing approach by the use of Deep Bidirectional LSTM networks (DBLSTM), by which we could increase the recognition accuracy for articulatory features.

1 Introduction

With more than 7,000 living languages [1] and many of them facing extinction [2, 3], preserving all languages in the world is next to impossible. Languages being spoken by only a few speakers are not of such a big social or economic interest that would be required for them to be preserved. There has been active research on the process of automatic language documentation, but with many language specific peculiarities, the process depends upon the knowledge and experience of linguists. We aim at supporting those human experts with natural language processing (NLP) technology. While NLP systems are readily available only for a few well-researched languages with a large speaker base, they are not available for under-resourced languages. Thus, the French-German ANR-DFG project *Breaking the Unwritten Language Barrier* (BULB) was initiated to develop technologies that would assist documentary linguists in documenting unknown and unwritten languages. BULB will build tools based on these technologies and validate them on three mostly unwritten African languages of the Bantu family: Basaa, Myene and Embosi [4].

Looking at the typical work flow for documenting unwritten languages, the starting point is the collection of audio data in the field. Based on these recordings, linguists derive the phonetic inventory. This process starts with segmenting the recorded speech into phoneme-like units and those segments need to be clustered based on phonetic similarity. As it is unknown which acoustic events do carry information, e.g., tones, linguists first have to decide on the sound inventory.

This paper is organized as follows: In the next Section, we provide an overview of relevant work in the field. We describe our approach for articulatory feature extraction in Section 3, followed by a description of our experimental setup in Section 4. The results are presented in Section 5. In the final Section 6, we conclude this paper with an outlook to future work.

*This work was realized in the framework of the ANR-DFG project BULB (STU 593/2-1 and ANR-14-CE35-002) and also supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

2 Related Work

2.1 Articulatory Feature Extraction

Articulatory features (AFs) represent the state of the articulators in the human vocal tract. Phones identify a certain configuration of articulatory features. Different approaches for using AFs in the field of speech recognition have been proposed. Metze and Waibel [5] proposed to use AFs in addition to the regular ASR pipeline in order to make systems more robust in terms of speaker or channel variability. It has also been shown that AFs are language independent and that these features can be recognized across languages [6, 7]. In addition to systems based on GMM/HMMs, AFs have also been used in ASR systems featuring neural networks [8].

2.2 Phoneme Discovery

Discovering phonemes in an unknown language is a difficult problem and subject of ongoing research. In the past, HMM based approaches were proposed[9]. Recent approaches published as part of the Zero Resource Speech Challenge [10] do make use of neural networks [11, 12], but there are GMM based methods [13, 14] as well. While detecting phones is difficult, clustering them into phonemes is a challenging task, even for linguists. According to Kempton and Moore [15], determining whether two phones are an allophone in a language is something which cannot be handled automatically because with languages having specific peculiarities, it is difficult to establish a language universal criteria.

2.3 Phoneme Segmentation

Prior to clustering phonemes, the recordings need to be segmented into single phoneme-like units. Early work in the field includes an HMM-based approach towards automatic segmentation and labeling of speech [16]. Current HMM-based models achieve boundary accuracies up to 96.8% [17] on certain tasks. In recent publications, we addressed the problem of automatically segmenting audio recordings into phoneme like units [18]. We demonstrated a first approach to derive the phoneme inventory of an unknown language using DBLSTMs.

3 Articulatory Feature Extraction and Phoneme Set Discovery

AFs describe the state of the articulators in the human vocal tract, representing the position of the tongue or opening of the mouth, for instance. In total, we used 7 different AF types, as shown in Table 1. The AFs can be divided into two categories: Consonants (cplace, ctype, cvox) and vowels (vfront, vheight, vlng, vrnd). As each type applies to only one category, we added an additional class representing “does not apply”. A certain configuration of the articulators represents a phone. With a limited phone inventory, only a limited set of AF configurations can be recognized. But recognizing AFs instead of phones mitigates this restriction and is therefore more universal. We extracted AFs using multilingually trained neural networks. Similar to previous works [19, 20], we used fully connected feed-forward deep neural networks (DNNs) to extract articulatory features. The networks used a context of 6 frames with 5 hidden layers and 1,000 neurons per layer. By the use of one network per feature, we could avoid co-adaptation where the networks might be biased by being exposed to only certain AF configurations during training.

In this work, we extended this approach by the use of Deep Bidirectional LSTM neural networks (DBLSTMs), which have shown to improve performance in tasks like speech recog-

Table 1 – Overview of AF types used

Type	# of Classes	Description
cplace	8	Place of articulation
ctype	6	Type of articulation
cvox	2	Voiced
vfront	3	Tongue x position
vheight	3	Tongue y position
vlng	4	Type of vowel
vrnd	2	Lips rounded

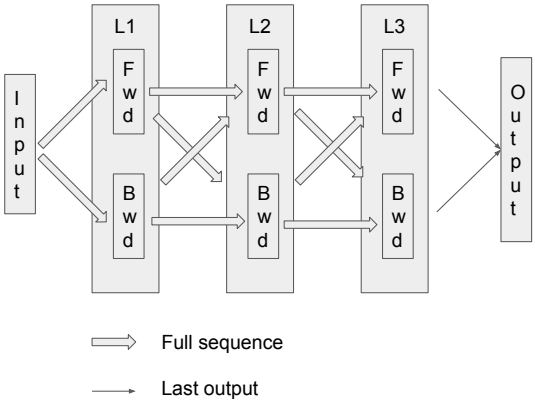

Figure 1 – DBLSTM network configuration: While using the output of the entire sequence throughout the recurrent layers, we only retain the final output for the output layer.

nition. The setup is inspired by [21]. We use 3 bidirectional LSTM layers with 500 cells per layer for each direction, without peephole connections. The network architecture is shown in Figure 1. We input a sequence of frames and use the full sequence output of the layers throughout the DBLSTM layers. After the last DBLSTM layer, we only use the final output of the network. This output is then fed into the output layer, a fully connected feed-forward layer with softmax activation.

4 Experimental Setup

We trained and evaluated neural networks on a combination of multiple languages, taken from the Euronews corpus [22]. This corpus consists of recordings from TV broadcast news with a total of 10 languages with matching acoustic conditions. Per language, 70h of data is available. We used the Janus Recognition Toolkit (JRTk) [23] which features the IBIS single-pass decoder [24].

Our setup for training the neural networks is based on Theano [25] and Lasagne [26]. By performing multiple experiments, we determined the optimal parameters for network training. Prior to training, we splitted the data into a training (90%) and validation set (10%). To extract

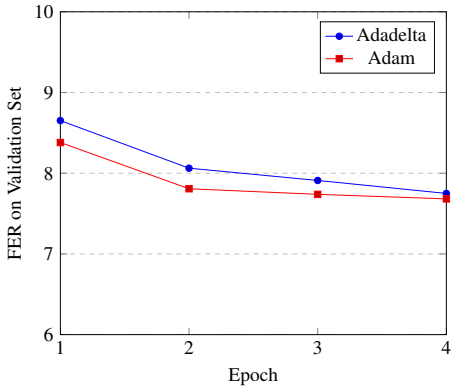

Figure 2 – Comparison of FER using Adam and Adadelta for updating the weights.

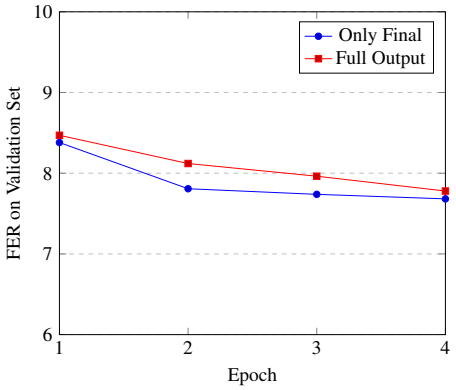

Figure 3 – DBLSTM Output Configurations: Using the entire or only the final output.

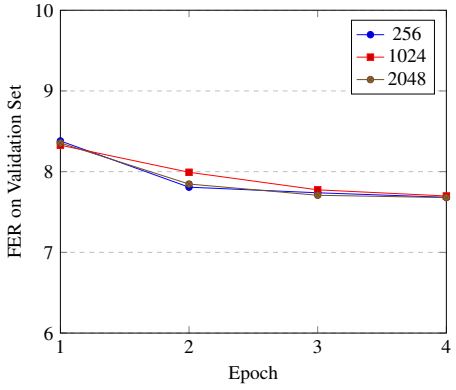

Figure 4 – Comparison of FER using mini-batches of size 256, 1024 and 2048.

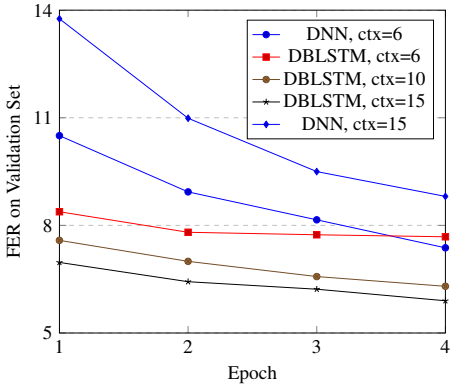

Figure 5 – FER of multiple context sizes, comparing DNNs and DBLSTMs

features from the audio for network training, we used a pre-processing based on a 32ms window with a frame-shift of 10ms. As input features, a combination of lMel and tonal features was used. Starting from a baseline configuration, we varied the mini-batch size as well as the size of the context. We also evaluated both Adam and Adadelta for computing the weight updates. Starting with the default learning rate, we updated the learning rate using an approach similar to new bob, where we decreased the learning rate by 0.5 once the error on the validation set increased. During these initial experiments, we limited the training to 4 epochs per configuration. After determining the optimal set of hyper parameters, we trained networks up to 8 epochs. For this set of experiments, we report on the frame error rate (FER) on the validation set.

5 Results

We evaluated several hyper parameter configurations. For this first series of experiments, we use only English monolingual data and only one articulatory feature *cplace*. This AF has 8+1 classes to distinguish. Based on the optimal configuration, we trained DBLSTM networks for each AF multi-lingually and compared the results to the DNN based baseline.

5.1 Parameter Updates

As baseline configuration, we chose to use a context of 6, batch size of 256, 3 layers and 500 neurons per layer. In this first experiment, we evaluated two methods for updating the parameters. After 4 epochs, the FER on the validation set using Adam is lower (7.7%) compared to Adadelta (7.8%), see Figure 2. But as the difference is only marginal, we continued to use Adam as the default update method.

5.2 LSTM Output Configuration

The outputs of DBLSTM networks can be used in different ways. One possibility is to feed the data into the network and retrieve the output for each frame in the entire sequence. Another possibility would be to retrieve only the output of the last sample. As shown in Figure 3, using the entire output of the LSTM layer (7.8% FER) or only the last one (7.7% FER) does not lead to large differences. Hence we decided to use only the outputs from the last sample in each sequence, as this would allow Theano to apply some additional optimizations that reduce training time.

5.3 Mini-batch Size

As next experiment, we varied the size of the mini-batches. Starting with 256, we increased the size to 1024 and 2048. Using larger mini-batches results in fewer updates of the network parameters. Increasing the size of the mini-batch updates leads to minimal changes in error rates with all three tested sizes resulting in 7.7% FER after 4 epochs. As shown in Figure 4, increasing the size does not affect the FER. We therefore decided to use 256 as the default mini-batch size.

5.4 Context

In the next experiment, we varied the size of the context, respectively the sequence length. Using different configurations, we started with the default context width of 6 frames from our DNN based approach. In addition, we increased the context to 10 and 15 frames. As contrastive experiment, we also trained a DNN using a context width of 15 and the same number of parameters as the DBLSTM. The results in Figure 5 show that DBLSTM based setups benefit from an increased context size, while DNNs do not. In total, we see an improvement from 7.2% FER to 5.4% FER by using DBLSTMs. Increasing the context did not lead to improvements for DNN based setups. As the accuracy kept improving with an increased context width, the next step would have been to evaluate context sizes beyond 15. But due to technical limitations, we had to postpone those experiments.

After determining the optimal hyper parameters for both DNN and DBLSTM based setups, we trained both types of networks over more epochs. Figure 6 show a comparison of both networks. The DBLSTM achieves a FER as low as 5.6%, while the DNN achieves 8.4%.

5.5 Multilingual Results

Based on the optimal configuration determined by the previous experiments, we trained networks for all 7 types of AFs. Instead of using only data from one language, we used a combination of three languages (German, French, Turkish) for this experiment. Using DBLSTMs, the FER improved cross the different AF types, with cplace showing the biggest improvements from 8.4% (DNN) to 5.7% (DBLSTM). The results for all AFs can be seen in Table 2.

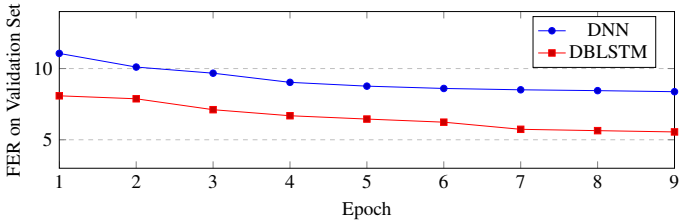

Figure 6 – FER of the best DNN (context = 6) and DBLSTM (context = 15) setup.

Table 2 – Classification error of AFs trained on German, French and Turkish using 70h per language. The results show the FER on the validation set.

Network Type	cplace	ctype	cvox	vfront	vheight	vlng	vrnd
DNN	8.4	8.2	7.8	7.2	7.9	7.3	6.1
DBLSTM	5.7	6.4	7.1	6.1	6.0	6.9	5.7
Relative Gain	33%	22%	9%	16%	25%	6%	7%

6 Conclusion

We have shown, that the AF recognition accuracy can be improved using DBLSTMs instead of DNNs. Future work includes the introduction of language adaptation techniques in order to further increase the performance of AF extraction cross lingually, as well a to estimate the amount of phoneme-like units.

References

[1] GRIMES, BARBARA F. AND PITTMAN, RICHARD SAUNDERS AND GRIMES, JOSEPH EVANS: *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, USA, 2005.

[2] NETTLE, D. and S. ROMAINE: *Vanishing Voices*. Oxford University Press Inc., New York, NY, USA, 2000.

[3] CRYSTAL, D.: *Language Death*. Cambridge University Press, Cambridge, UK, 2000.

[4] STÜKER, S., G. ADDA, M. ADDA-DECKER, O. AMBOUROUE, L. BESACIER, D. BLACHON, H. BONNEAU-MAYNARD, P. GODARD, F. HAMLAOUI, D. IDIATOV, G.-N. KOUARATA, L. LAMEL, E.-M. MAKASSO, M. MÜLLER, A. RIALLAND, M. V. DE VELDE, F. YVON, and S. ZERBIAN: *Innovative Technologies for Under-Resourced Language Documentation: The BULB Project*. In *2nd Workshop Collaboration and Computing for Under-Resourced Languages (CCURL 2016)*. 2016.

[5] METZE, F. and A. WAIBEL: *A Flexible Stream Architecture for ASR Using Articulatory Features*. In *INTERSPEECH*. 2002.

[6] STÜKER, S., T. SCHULTZ, F. METZE, and A. WAIBEL: *Multilingual Articulatory Features*. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol. 1, pp. 144–147. IEEE, Hong Kong, 2003.

[7] STÜKER, S., F. METZE, T. SCHULTZ, and A. WAIBEL: *Integrating Multilingual Articulatory Features into Speech Recognition*. In *Proceedings of the 8th European Conference*

on Speech Communication and Technology EUROSPEECH'03, pp. 1033–1036. ISCA, Geneve, Switzerland, 2003.

[8] SWIETOJANSKI, P., A. GHOSHAL, and S. RENALS: *Unsupervised Cross-Lingual Knowledge Transfer in DNN-based LVCSR*. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pp. 246–251. IEEE, 2012.

[9] VARADARAJAN, B., S. KHUDANPUR, and E. DUPOUX: *Unsupervised learning of acoustic sub-word units*. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 165–168. Association for Computational Linguistics, 2008.

[10] VERSTEEGH, M., R. THIOLLIERE, T. SCHATZ, X. N. CAO, X. ANGUERA, A. JANSEN, and E. DUPOUX: *The Zero Resource Speech Challenge 2015*. In *Proceedings of Interspeech*. 2015.

[11] RENSHAW, D., H. KAMPER, A. JANSEN, and S. GOLDWATER: *A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge*. In *Proceedings of Interspeech*. 2015.

[12] BADINO, L., A. MERETA, and L. ROSASCO: *Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders*. In *Proceedings of Interspeech*. 2015.

[13] CHEN, H., C.-C. LEUNG, L. XIE, B. MA, and H. LI: *Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling: A feasibility study*. In *Proceedings of Interspeech*. 2015.

[14] HECK, M., S. SAKTI, and S. NAKAMURA: *Unsupervised linear discriminant analysis for supporting dpgmm clustering in the zero resource scenario*. *Procedia Computer Science*, 81, pp. 73–79, 2016.

[15] KEMPTON, T. and R. K. MOORE: *Discovering the Phoneme Inventory of an Unwritten Language: A Machine-Assisted Approach*. *Speech Communication*, 56, pp. 152–166, 2014.

[16] BRUGNARA, F., D. FALAVIGNA, and M. OMOLOGO: *Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models*. *Speech Communication*, 12(4), pp. 357–370, 1993.

[17] STOLCKE, A., N. RYANT, V. MITRA, J. YUAN, W. WANG, and M. LIBERMAN: *Highly Accurate Phonetic Segmentation Using Boundary Correction Models and System Fusion*. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 5552–5556. IEEE, 2014.

[18] FRANKE, J., M. MÜLLER, S. STÜKER, and A. WAIBEL: *Phoneme boundary detection using deep bidirectional lstms*. In *Speech Communication; 12. ITG Symposium; Proceedings of*. VDE, 2016.

[19] MÜLLER, M., S. STÜKER, and A. WAIBEL: *Using Language Adaptive Deep Neural Networks for Improved Multilingual Speech Recognition*. *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*, 2016.

[20] MÜLLER, M., J. FRANKE, S. STÜKER, and A. WAIBEL: *Towards Phoneme Inventory Discovery for Documentation of Unwritten Languages*. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017.

[21] ZEYER, A., R. SCHLÜTER, and H. NEY: *Towards Online-Recognition with Deep Bidirectional LSTM Acoustic Models*. In *Proceedings of the Interspeech*. San Francisco, CA, USA, 2016.

[22] GRETTER, R.: *Euronews: A Multilingual Benchmark for ASR and LID*. In *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.

[23] WOSZCZYNA, M., N. AOKI-WAIBEL, F. D. BUØ, N. COCCARO, K. HORIGUCHI, T. KEMP, A. LAVIE, A. MCNAIR, T. POLZIN, I. ROGINA, C. ROSE, T. SCHULTZ, B. SUHM, M. TOMITA, and A. WAIBEL: *JANUS 93: Towards Spontaneous Speech Translation*. In *International Conference on Acoustics, Speech, and Signal Processing 1994*. Adelaide, Australia, 1994.

[24] SOLTAU, H., F. METZE, C. FUGEN, and A. WAIBEL: *A One-Pass Decoder Based on Polymorphic Linguistic Context Assignment*. In *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pp. 214–217. IEEE, 2001.

[25] THEANO DEVELOPMENT TEAM: *Theano: A Python Framework for Fast Computation of Mathematical Expressions*. *arXiv e-prints*, abs/1605.02688, 2016. URL `http://arxiv.org/abs/1605.02688`.

[26] DIELEMAN, S., J. SCHLÜTER, C. RAFFEL, E. OLSON, S. K. SØNDERBY, D. NOURI, D. MATURANA, M. THOMA, E. BATTENBERG, J. KELLY, J. D. FAUW, M. HEILMAN, DIOGO149, B. MCFEE, H. WEIDEMAN, TAKACSG84, PETERDERIVAZ, JON, INSTAGIBBS, D. K. RASUL, CONGLIU, BRITEFURY, and J. DEGRAVE: *Lasagne: First release.* 2015. doi:10.5281/zenodo.27878. URL `http://dx.doi.org/10.5281/zenodo.27878`.