TOWARDS PHONEME INVENTORY DISCOVERY FOR DOCUMENTATION OF UNWRITTEN LANGUAGES

Markus Müller, Jörg Franke, Alex Waibel, Sebastian Stüker

Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
Email: {m.mueller,alexander.waibel,sebastian.stueker}@kit.edu
joerg.franke@student.kit.edu
Web: isl.anthropomatik.kit.edu

ABSTRACT

Documenting unwritten languages is a challenging task, even for trained specialists. To help linguists in better and faster documenting new languages is the goal of the French-German ANR-DFG project BULB. To discover the phonetic inventory of a language the project follows three steps: estimating phoneme boundaries, classifying articulatory features (AFs) for each individual segment and clustering the segments into a phoneme inventory. In this work, we focus on estimating the phoneme boundaries and the extraction of AFs, but also perform a first simple clustering based on the recognized AFs. We demonstrate that our Deep Bidirectional LSTM-based approach for identifying phoneme boundaries achieves state-of-the-art performance and evaluate AF extraction based on feed forward neural networks.

Index Terms— Articulatory Features, DBLSTMs, Multilingual, Phoneme Segmentation, Language Documentation

1. INTRODUCTION

There are 7,000 living languages in the world. With the exception of a few well researched and ressourced languages, there exists a long tail of languages that are not documented, only spoken by a few speakers and in danger of becoming extinct [1]. Documentation is required in order to preserve the cultural heritage of these languages. With the vast majority of these languages being unwritten, the documentation process is even more time consuming. Utilizing Natural Language Processing (NLP) systems could help in accelerating this process.

For this reason, the French-German ANR-DFG project *Breaking the Unwritten Language Barrier* (BULB) was initiated to develop technologies that would assist documentary linguists in charting unknown and unwritten languages. BULB aims at building tools based on these technologies and validating them on three mostly unwritten African languages of the Bantu family: Basaa, Myene and Embosi [2]. Recently, we proposed methods for automatic phoneme segmentation [3, 4]. In this work, we now combine these approaches with articulatory feature detection and evaluate this setup using data from Basaa.

Our paper is organized as follows: In the next section, we provide an overview of related work in the field. Section 3 describes our approach to phoneme segmentation and section 4 outlines the extraction of articulatory features (AFs). The experimental setup is presented in section 5, followed by the results in section 6. This paper concludes with section 7 where we also provide an outlook on 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. Phoneme Segmentation

Segmenting recordings into single phonemes is a first step in discovering the phoneme set of an unknown language prior to documenting it. Methods for scenarios where no data from the target language is available are being explored in the Zero Speech Challenge [5]. One approach is to use a speech recognition system to recognize phonemes [6] and then discard their identity, only retaining the phoneme boundaries. Since such a system usually has a limited phoneme inventory, dealing with unknown languages that introduce new phonemes may prove difficult. While [6] only used an English ASR system for detecting phoneme boundaries this way, [3] evaluated the effect of using different monolingual and multilingual systems. Recent work using neural networks for directly classifying phoneme boundaries, instead of recognizing phoneme sequences, achieved even superior performance [7, 4].

2.2. Articulatory Feature Extraction

Articulatory features represent the target of the articulators in the vocal tract when pronouncing a specific phone. The combination of AFs determines the identity of that specific phone spoken, i.e., phones are just a short-hand for bundles

of AFs. The use of articulatory features for speech recognition has been proposed in the past: [8] proposed the use of AFs as additional detectors for ASR. [9, 10] has shown that articulatory features can be recognized more robustly across languages than phonemes. The authors have shown that the phoneme coverage of multilingual AF recognizers on a new language in general is also larger compared to multilingual phoneme recognizers. Neural network based setups also benefit from AFs [11, 12].

2.3. Language Feature Vectors

Documenting unknown languages implies the lack of transcribed recordings. Since we want to perform phoneme segmentation and AF detection on these languages, techniques handling such cross-lingual scenarios are required. In the field of automatic speech recognition (ASR), there exist methods like i-Vectors or bottleneck speaker vectors (BSVs) [13] to adapt neural networks to different speakers. These methods show that neural networks benefit from additional input modalities. In the notion of BSVs, we have shown that this principle can also be applied in a multi- or cross-lingual scenario. While providing the language identity information (LID) alone leads to improvements [14], using language feature vectors (LFVs) instead leads to even bigger gains [15, 16]. The advantage of LFVs compared to LID is that they are applicable to unseen languages as well.

2.4. Phoneme Discovery

The unsupervised discovery of linguistic units is the topic of ongoing research. There exist HMM based approaches like [17]. In the context of the Zero Resource Speech Challenge [5], there are more recent approaches using neural networks [18], but also GMM based methods [19].

3. PHONEME SEGMENTATION

3.1. LVCSR based

One way to segment audio into phoneme-like units is to use phoneme recognition. For this we used the Janus Recognition Toolkit (JRTk) [20] which features the IBIS single-pass decoder [21]. We modified a multilingually trained system for large vocabulary continuous speech recognition (LVCSR) to recognize individual phonemes [3]. The boundaries of these hypothesized phonemes were retained while the phoneme labels were discarded [6].

3.2. DBLSTM based systems

In addition to LVCSR based systems, we also evaluated a different approach based on neural networks [4]. Using a standard pre-processing pipeline, we extracted 40 dimensional lMel and 14 dimensional tonal features to train a deep bi-directional LSTM (DBLSTM) network with two layers to detect phoneme boundaries. The first layer contained 300 LSTM-nodes and the second layer 100 nodes, each node featured peephole connections. We used back propagation through time with mini batch stochastic gradient decent updates. The optimization was performed using AdaDelta, an extension to AdaGrad with the advantage of having a smoother gradient adaption with no need for manual tuning of the learning rate. The ratio of frames marked as "boundary" to "no boundary" was 1:8. In order to deal with this imbalance in the class distribution, we increased the weight of the "boundary" loss.

4. ARTICULATORY FEATURE EXTRACTION

To infer the phoneme inventory of an unknown language, the first step is to extract articulatory features (AFs) for each phoneme-like segment. Based on JRTk, we built an LVCSR system to generate phoneme labels for the recordings using a resolution of 10ms. The training of the LVCSR system required a pronunciation dictionary, which we created using MaryTTS. HMM-based ASR systems typically model each phoneme using 3 states (begin, middle, end) to account for co-articulation. Previous work has shown, that using only the middle frames for training AF detectors leads to the best results [22].

With the AF definitions embedded in MaryTTS, we established a mapping from phonemes to AFs. We used 7 different types of AFs, as shown in Table 1, with each type having different targets, e.g. "ctype" has 6 targets: Stop, fricative, affricative, liquid, nasal and approximant. The types fall into two categories: AFs for vowels (with prefix *v*) and consonants (with prefix *c*). As each type only applies to one category, we added an additional class that represented "does not apply". We trained feed forward neural networks for estimating AFs.

Type	**# 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

Table 1. Overview of AF types used

The neural network architecture used is based on LVCSR systems, featuring 5 hidden layers with 1,600 neurons each. To pre-process the audio, we used our standard pipeline with a frame-size of 32ms and a frame-shift of 10ms to extract features. Using a context of +/- 6 frames, we fed these features into the network. To prevent co-adaptation between language

specific combinations of AFs, we trained individual networks for each AF.

5. EXPERIMENTAL SETUP

5.1. Corpora

We used training data from the Euronews corpus [23] for our experiments. It consists of recordings from TV broadcast news in 10 different languages, with 70h of data per language. Depending on individual experiments, we only used a subset of these languages. In addition, we used 2h of Basaa recordings. This data set contains of recordings of utterances that were re-spoken by a single speaker in a clean environment. For further details about this data set, please refer to [3].

5.2. Phoneme Boundary Detection

The systems for phoneme boundary detection were trained using data from 5 languages (**FR**ench, **GE**rman, **IT**alian, **RU**ssian, **TU**rkish). We selected these languages based on the availability of pronunciations from MaryTTS. To evaluate the system performance, we used English data. For neural network training, the data from each language was divided into two sets: a training set containing 90% of the data and a validation set containing 10%. To generate reference boundaries, we used an LVCSR system to force align the transcripts to the audio. For evaluation of the hypothesized boundaries, we used precision, recall and F1 score. In order to determine whether a predicted boundary matches a boundary in the reference, we allow for a margin of error of 20ms, which is common in literature [24].

5.3. Articulatory Feature Extraction

To train and evaluate AF extraction, we narrowed the set of languages to 4 (EN, FR, GE, TR) because the phoneme and AF definitions for Russian and Turkish in MaryTTS differed to some extend from those of the other 4 languages.

6. RESULTS

6.1. Phoneme Segmentation

We evaluated our approach for phoneme segmentation using two different conditions. Both setups were trained on the same set of data from Euronews (5 languages: FR, GE, IT, RU, TR). For the first evaluation, we detected phoneme boundaries using English in-domain data from Euronews with matching acoustic conditions (see Table 2, left columns). The results for *GMM*, *DNN* and *LFV* were taken from [3, 15] and are provided for reference. *GMM* corresponds to a context-independent LVCSR system based on GMM/HMMs. *DNN* represents a context-dependent LVCSR system with a hybrid DNN/HMM acoustic model and Deep Belief Network Features (DBNF) based pre-processing. The setup of the system labelled *LFV* is identical to *DNN*, but with the addition of LFVs to the acoustic input features [15]. The *DBLSTM* system corresponds to the setup described in Section 3.2. It produced the results with the highest score, outperforming all LVCSR based systems. A possible reason for this is that LVCSR systems are being trained to recognize the most likely word sequence, but not individual words or phonemes at precise points in time. The DBLSTM on the other hand was trained to recognize phoneme boundaries at precise points in time. Using the same setups, we also estimated phoneme boundaries on Basaa data. Although the scores (Table 2, right columns) are lower compared to our segmentation on English data, this experiment proves that our setup is also suited for languages other than English. The lower F-score (0.68 vs. 0.79) could be explained by different acoustic conditions, as the Basaa data was re-spoken in a quiet room in contrast to TV broadcast news which quite frequently feature ambient noises like background music. In order to minimize these differences, training with data covering multiple conditions is necessary.

System	Precision		Recall		F-Score	
	EN	BAS	EN	BAS	EN	BAS
GMM	0.63	0.47	0.67	0.54	0.65	0.50
DNN	0.65	0.52	0.70	0.52	0.67	0.52
LFV	0.67	0.54	0.73	0.53	0.70	0.54
DBLSTM	**0.74**	**0.68**	**0.84**	**0.72**	**0.79**	**0.69**

Table 2. Results for cross-lingual phoneme segmentation on English (**EN** left columns) and Basaa (**BAS** right columns).

6.2. AF extraction cross-/multilingual

Setup	cplace	ctype	cvox	vfront	vrnd
BL 3L	8.37	8.18	7.79	7.16	6.15
EN CL	15.93	15.60	14.29	14.07	9.49
BL 4L	8.56	8.42	8.10	7.50	6.01
EN ML	8.34	8.01	8.22	7.71	5.23

Table 3. Classification error of AFs using 70h of GE, FR, TR (3L) or 70h of GE, EN, FR, TR (4L). Evaluation on training languages (*BL 3L* and *BL 4L*) or cross- (CL) / multilingually (ML) on English. A selection of AFs is shown.

We evaluated our setup by training networks for AF detection using 70h of data from 3 languages (GE, FR, TR) and testing cross-lingual on English (*EN CL*). The results for a subset of AFs are shown in the upper part of Table 3. For reference, we included the classification error of a multilingual

system trained on all 4 languages (*Baseline 4L*) as well as the classification error of this system only on English (*EN ML*). In the cross-lingual case, the error increased throughout different AFs, which can be explained by the AF mapping established by MaryTTS not being completely language independent. There exist some minor differences between languages that may account for this increase. As shown in Section 6.3, the recognition quality of the AFs is still good enough for detecting phonemes of the target language. Further techniques in both data normalization and cross-lingual adaptation are required to increase the recognition accuracy.

6.3. Phoneme Discovery

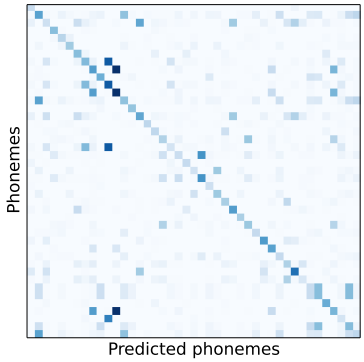

Fig. 1. Multilingual phoneme mapping: Mapping AFs to English phoneme targets. System was trained on GE, EN, FR and TR.

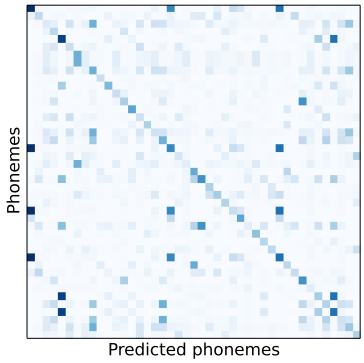

Fig. 2. Crosslingual phoneme mapping: Mapping AFs to English phoneme targets based on detected phoneme boundaries. System was trained on GE, FR and TR.

In order to discover the phoneme inventory of an unknown language, we combined both techniques. First, we segmented the audio into phoneme-like units using the DBLSTM based setup. Second, we extracted AFs based on these segments. By stacking the outputs of each AF network, we generated a 36 dimensional feature vector. Since we do not have a sub-phoneme state-level alignment to divide the frames into begin, middle and end states, we approximated this by averaging the features over the inner third of frames for each phoneme.

We clustered the per segment AFs using KMeans clustering. This method requires the number of classes as parameter. For this work, we assumed the class count to be known for the target language, enabling us to assess how good the phoneme set could be reconstructed given the amount of phonemes. Additional research is required to determine the number of classes automatically.

We compared two setups: As baseline, we used a multilingual system, trained using data from 4 languages (GE, EN, FR, TR) and evaluated the mapping on English. As shown in Figure 1, the mapping of the inferred phonemes to actual phonemes can be achieved to a great extend, although there are a few classes that produce ambiguous matches. The system we are comparing against was trained on 3 languages (GE, FR, TR) to extract AFs. Figure 2 shows the result after clustering and mapping of the phonemes. There are a few more outliers introducing more ambiguity, but a mapping similar to our multilingual approach could be established.

This experiment demonstrated, that by combination of phoneme segmentation and AF extraction a phoneme set can be discovered. With this method, it is possible to support linguists in the documentation of unwritten languages. Using an iterative approach with the human in the loop, it is possible fine-tune the amount of hypothesized phonemes by manually examining mappings with low confidence scores.

7. CONCLUSION AND OUTLOOK

In this work, we evaluated methods aimed at discovering phoneme inventories of unwritten languages. First, we addressed the problem of segmenting recordings into phoneme-like units. Using a DBLSTM based setup resulted in the best classification performance. Second, we clustered these phoneme-like units based on AFs. As shown in the confusion matrices, the resulting classes correspond to a great extent to the actual phonemes present in the target language. Future work includes improving the performance of AFs to achieve a better and more accurate phoneme clustering.

8. REFERENCES

[1] Daniel Nettle and Suzanne Romaine, *Vanishing Voices*, Oxford University Press Inc., New York, USA, 2000.

[2] Sebastian Stüker, Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Markus Müller, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian, "Innovative Technologies for Under-Resourced Language Documentation: The BULB

Project," in *2nd Workshop Collaboration and Computing for Under-Resourced Languages (CCURL 2016)*, 2016.

[3] Marco Vetter, Markus Müller, Fatima Hamlaoui, Graham Neubig, Satoshi Nakamura, Sebastian Stüker, and Alex Waibel, "Unsupervised Phoneme Segmentation of Previously Unseen Languages," in *Proceedings of the Interspeech*, 2016.

[4] Jörg Franke, Markus Müller, Sebastian Stüker, and Alex Waibel, "Phoneme Boundary Detection using Deep Bidirectional LSTMs," in *Speech Communication; 12. ITG Symposium; Proceedings of*. VDE, 2016.

[5] Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux, "The Zero Resource Speech Challenge 2015," in *Proceedings of Interspeech*, 2015.

[6] Prasanna Kumar Muthukumar and Alan W. Black, "Automatic Discovery of a Phonetic Inventory for Unwritten Languages for Statistical Speech Synthesis," in *ICASSP, 2014 IEEE International Conference on*. IEEE, 2014, pp. 2613–2617.

[7] Sandrine Mouysset Thomas Pellegrini, "Inferring Phonemic Classes from CNN Activation Maps Using Clustering Techniques," in *INTERSPEECH*, San Francisco, USA, September 2016.

[8] Florian Metze and Alex Waibel, "A Flexible Stream Architecture for ASR Using Articulatory Features," in *INTERSPEECH*, 2002.

[9] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual Articulatory Features," in *ICASSP*. 2003, vol. 1, pp. 144–147, IEEE.

[10] S. Stüker, F. Metze, T. Schultz, and A. Waibel, "Integrating Multilingual Articulatory Features into Speech Recognition," in *EUROSPEECH*, Geneve, Switzerland, 2003, pp. 1033–1036, ISCA.

[11] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Unsupervised Cross-Lingual Knowledge Transfer in DNN-based LVCSR," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 246–251.

[12] Markus Müller, Sebastian Stüker, and Alex Waibel, "Towards Improving Low-Resource Speech Recognition Using Articulatory and Language Features," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, U.S.A., 2016.

[13] Hengguan Huang and Khe Chai Sim, "An Investigation of Augmenting Speaker Representations to Improve Speaker Normalisation for DNN-based Speech Recognition," in *ICASSP*. IEEE, 2015, pp. 4610–4613.

[14] Markus Müller and Alex Waibel, "Using Language Adaptive Deep Neural Networks for Improved Multilingual Speech Recognition," *IWSLT*, 2015.

[15] Markus Müller, Sebastian Stüker, and Alex Waibel, "Language Adaptive DNNs for Improved Low Resource Speech Recognition," in *Interspeech*, 2016.

[16] Markus Müller, Sebastian Stüker, and Alex Waibel, "Language Feature Vectors for Resource Constraint Speech Recognition," in *Speech Communication; 12. ITG Symposium; Proceedings of*. VDE, 2016.

[17] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel 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*. Association for Computational Linguistics, 2008, pp. 165–168.

[18] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater, "A Comparison of Neural Network Methods for Unsupervised Representation Learning on the Zero Resource Speech Challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] Michael Heck, Sakriani Sakti, and Satoshi Nakamura, "Unsupervised Linear Discriminant Analysis for Supporting DPGMM Clustering in the Zero Resource Scenario," *Procedia Computer Science*, vol. 81, pp. 73–79, 2016.

[20] Monika Woszczyna et al., "JANUS 93: Towards Spontaneous Speech Translation," in *International Conference on Acoustics, Speech, and Signal Processing 1994*, Adelaide, Australia, 1994.

[21] Hagen Soltau, Florian Metze, Christian Fugen, and Alex Waibel, "A One-Pass Decoder Based on Polymorphic Linguistic Context Assignment," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 214–217.

[22] Florian Metze, *Articulatory Features for Conversational Speech Recognition*, Ph.D. thesis, Karlsruhe, Univ., Diss., 2005, 2005.

[23] Roberto Gretter, "Euronews: A Multilingual Benchmark for ASR and LID," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[24] Odette Scharenborg, Vincent Wan, and Mirjam Ernestus, "Unsupervised Speech Segmentation: An Analysis of the Hypothesized Phone Boundaries," *Acoustical Society of America, Journal of*, vol. 127, no. 2, pp. 1084–1095, 2009.