INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Unsupervised Phoneme Segmentation of Previously Unseen Languages

Marco Vetter[1], Markus Müller[1], Fatima Hamlaoui[2] Graham Neubig[3], Satoshi Nakamura[3]
Sebastian Stüker[1], Alex Waibel[1]

[1]Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany
[2]Zentrum für Allgemeine Sprachwissenschaft, Berlin, Germany
[3]Augmented Human Communications Laboratory, Nara Institute of Science and Technology, Japan

ucbqi@student.kit.edu, {m.mueller|sebastian.stueker|alexander.waibel}@kit.edu
hamlaoui@zas.gwz-berlin.de, {neubig|s-nakamura}@is.naist.jp

Abstract

In this paper we investigate the automatic detection of phoneme boundaries in audio recordings of an unknown language. This work is motivated by the needs of the project BULB which aims to support linguists in documenting unwritten languages. The automatic phonemic transcription of recordings of the unwritten language is part of this. We cannot use multilingual phoneme recognizers as their phoneme inventory might not completely cover that of the new language. Thus we opted for pursuing a two step approach which is inspired by work from speech synthesis for previously unknown languages. First, we detect boundaries for phonemes, and then we classify the detected segments into phoneme units. In this paper we address the first step, i.e. the detection of the phoneme boundaries. For this we again used multilingual and crosslingual phoneme recognizers but were only interested in the phoneme boundaries detected by them and not the phoneme identities. We measured the quality of the segmentations obtained this way using precision, recall and F-measure. We compared the performance of different configurations of mono- and multilingual phoneme recognizers among each other and against a monolingual gold standard. Finally we applied the technique to Basaa, a Bantu language.

Index Terms: Automatic phoneme transcription, multilingual speech recognition, language documentation

1. Introduction

There currently exist over 7,000 living languages in the world [1]. A large number of these are only spoken by a small group of speakers and are being threatened by extinction [2, 3]. While *Natural Language Processing* (NLP) systems have been successfully built for many languages with a large speaker base or great economic value, they are not available for the vast majority of smaller, under-resourced languages. The need for extensive, annotated training corpora usually makes building such systems costly and time-consuming. Additionally, many small languages and regional dialects of major languages do not feature a standardized writing system, complicating the creation of NLP systems for them.

The number of endangered languages is so large that their comprehensive documentation by the community of documentary linguists will only be possible if supported by NLP technology. Therefore it is the goal of the French-German ANR-DFG project *Breaking the Unwritten Language Barrier* (BULB) to develop tools to enable the efficient automatic processing of unwritten languages. Initial targets will be three mostly unwritten African languages of the Bantu family (Basaa, Myene and Embosi) [4].

One of BULB's goals is to automatically segment recordings of new languages into phonemes. As no prior knowledge of the target language is available, the use of multilingual phoneme recognizers for this task is not possible, since their phoneme inventory might not sufficiently cover the target phoneme inventory. We therefore decided to pursue the two step approach of first detecting phoneme boundaries, followed by classifying the detected segments into phonemes.

In this paper we address the first step, i.e. phoneme segmentation, by using crosslingual and multilingual phoneme recognizers. For this we will exclusively focus on the positions of detected phoneme boundaries, disregarding the identity of the phonemes detected. We compare the performance of different cross- and multilingual phoneme recognizers among each other and, where available, against a monolingual gold standard consisting of a phoneme recognizer trained on the target language in a traditional supervised manner.

2. Related Work

Significant work has been done on the topic of building speech recognition systems for unwritten and under-resourced languages. [5] tried to estimate phoneme boundaries by analyzing the acoustic change of audio signals. They proposed a two step method where the information derived from the speech signal is expanded by additional cues. [6] presents an approach to discovering a proper set of subword-like units. In addition to segmenting the audio, they also train a Dirichlet process mixture model for representing individual acoustic units. [7] has investigated algorithms and metrics for the task of unsupervised phoneme segmentation. In [8] the authors presented an HMM/SVM approach for automatic phoneme segmentation that imitates the human phoneme segmentation process.

Recent work has been done in the context of the Zero Speech Challenge [9]. This challenge focuses on the unsupervised discovery of subword units from raw speech. The organizers provide a unified and open suite of evaluation metrics.

3. Unsupervised Segmentation

Our approach to unsupervised segmentation of speech recordings into phonemes in a new, unknown language is inspired by work for speech synthesis in [10, 11]. In this work an English phoneme recognizer was used to segment various languages. However, the authors did not evaluate the quality of the seg-

http://dx.doi.org/10.21437/Interspeech.2016-1440

mentation directly. Instead they only considered the extrinsic quality of the resulting speech synthesis system. For our task, i.e. the accurate phonetic transcription of a language for documentation purposes, the quality of the resulting segmentation is much more important than for the task of finding segments suitable for speech synthesis. We therefore carefully measure the quality of the segmentation of speech into phonemes using F-measure, disregarding the identity of the recognized phonemes.

Furthermore, [10, 11] only used an English monolingual phoneme recognizer. However from previous experience we know that when working across languages, multilingual acoustic models usually outperform monolingual ones. We therefore extend the approach to using multilingual acoustic models whose modeling units have been trained on multiple languages using a common phoneme set [12].

4. Experimental Setup

The recognizers for the experiments presented in this paper were built using the Janus Recognition Toolkit (JRTk) [13], which features the IBIS single-pass decoder [14]. For the creation of the pronunciation dictionaries we used MaryTTS [15].

A significant parameter used for decoding the target audio is the word penalty (LP) which is added to the score of a hypothesis for every word in the hypothesis. Higher LP values result in fewer hypothesized boundaries, and vice versa. Comparing the resulting number of segments to the number of segments to be expected according to the reference we can calculate an over- or under-segmentation ratio as defined in [5].

4.1. Basaa data

Basaa is one of the three Bantu languages of the BULB project. It is spoken by approximately 300 000 speakers (SIL 2005, [16]) from the Centre and Littoral regions in southern Cameroon.

The Basaa data used in the present experiment consists of ≈2 hours of re-spoken radio broadcasts. The original audio files were obtained from the radio station CRTV-Centre and feature a male native speaker of the language. His speech was phonetically transcribed by a linguist and later carefully re-spoken by a female native speaker of Basaa in a quiet environment.

4.2. English Data

For the sake of comparison, considering the absence of a usable baseline system for Basaa, we also used English as a *faux* unseen target language. This allows us to automatically create labels to use as a ground truth, even when using data that does not feature phoneme-level annotations. For this purpose the target data's orthographic annotations were converted to a phonetic representation via the G2P component of MARY [15] and then matched to the audio via a forced alignment performed by a pre-trained English language ASR system.

As source languages we chose German, French, Italian, Russian and Turkish. Both target and source audio were taken from the Euronews Corpus [17]. Euronews data is composed of news recordings originally broadcast on the channel of the same name. Recordings consist of read and planned speech, often spoken over news footage with separate audio running in the background, making the collected data noisy.

Individual source language systems were trained on 68-77 hours of training data each. Since Euronews does not provide manual annotations on a phonetic level, we also trained a recognizer on the same amount of English Euronews training audio in order to create phoneme annotations against which to compare our cross- and multilingual segmentations. The multilingual system (denoted as M5 throughout this paper) was trained on a combined 360 hours of audio, with equal parts taken from each of the five source languages. For details on the distribution of the source language training data also see Table 1. For testing, we used two sets of English audio files: a longer one with ≈29minutes of speech across 29 news reports, and a shorter one, consisting of ≈24minutes of speech across a subset of 28 news reports.

Language	Length	#Phonemes	Cov. EN	Cov. BA
EN	72.8h	40	–	–
DE	73.2h	56	82.5%	80.6%
FR	68.1h	33	57.8%	61.3%
IT	77.2h	59	60.0%	74.2%
TR	70.4h	26	55.0%	54.8%
RU	72.2h	25	37.5%	45.2%
M5	361.3h	99	85.0%	96.8%

Table 1: Amount of audio data used for training mono- and multilingual recognizers (in hours), number of phonemes used in training and phoneme coverage on English and Basaa

It should be noted that the individual systems trained on the data for the source languages listed above feature vastly different phonetic coverage on the target language with regard to their acoustic model, as presented in Table A baseline for our experiments on English will be provided via segmentations generated by an ASR system trained on the English Euronews training data. To measure performance on clean speech we also used English speech taken from the TIMIT corpus as target audio. TIMIT [18] provides recordings from 630 native speakers of eight dialects of American English recorded in a controlled studio environment. The audio is therefore quite clean, with no background noise whatsoever, and thus differs significantly in nature from the Euronews data used for training the systems.

5. Experiments and Results

Before discussing results, section 5.1 will briefly introduce the employed metrics and how they were applied. Initial experiments on English were conducted without a language model; the results are presented in sections 5.2 (for noisy speech) and 5.3 (for clean speech). In addition, we ran experiments with special language models estimated on the phoneme sequences of the training data in order to see if this information is beneficial, and compared the results to those obtained without language models (section 5.4). Finally, we attempted to segment Basaa speech with both mono- and multilingual systems (section 5.5).

5.1. Performance metrics

To evaluate the accuracy of the phoneme segmentation generated by our systems we use precision, recall and F1 scores of the detected phoneme boundaries. In previous work on the subject of segmentation, such as [5], [6] and [7], researchers have opted to allow for a certain inaccuracy when determining whether a predicted boundary matches the reference. This is reasonable, as requiring matches to be exact to a single frame is a very strict standard that is hard to meet and does not account for the ambiguous nature of speech signals. A commonly chosen value

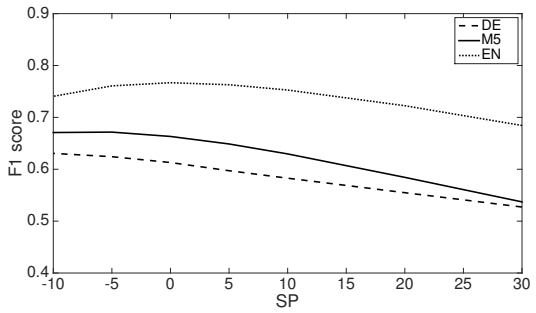

Figure 1: F1 scores for different silence penalties on English Euronews audio (no language model)

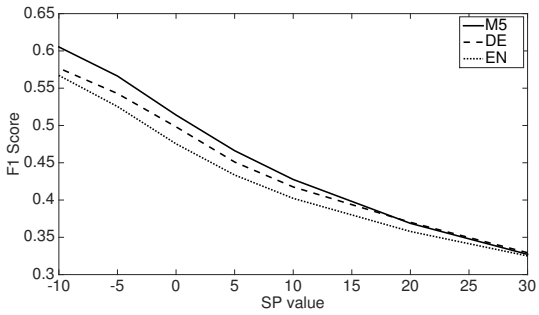

Figure 2: F1 scores for different silence penalties on English TIMIT audio (no language model)

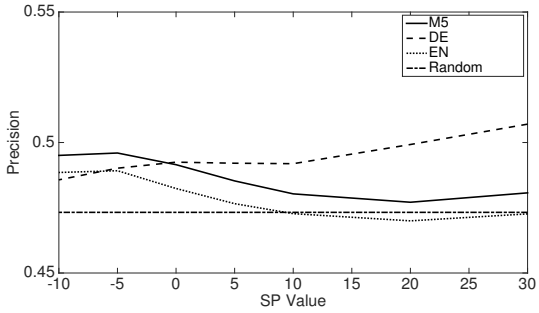

Figure 3: Precision scores for different silence penalties on English TIMIT audio (no language model)

for this *tolerance* is 20 ms, as larger windows quickly escalate scores and therefore are not particularly useful for evaluation.

5.2. Results on noisy English speech

During the experiments presented in this section and in section 5.3, we tagged all words (i.e. phonemes) as noise models. Hence we ran the experiments by varying the silence penalty (SP), which in this scenario corresponds to changes of LP in the second set of experiments where we introduced language models. In our experiments the number of segments produced was most accurate (i.e. resulted in minimal over- or undersegmentation) for a penalty value of 0. This is also where the baseline segmentation performed best according to the F1 score, as can be seen in Figure 1.

System	Precision	Recall	F-Score
EN (baseline)	0.7623	0.7715	0.7669
M5	**0.6299**	**0.6984**	**0.6624**
DE	0.5881	0.6440	0.6130
FR	0.6706	0.6791	0.6748
IT	0.6286	0.7016	0.6631
RU	0.5559	0.7372	0.6338
TR	0.6208	0.6709	0.6449

Table 2: Comparison of segmentation quality on English Euronews audio without language models

All monolingual systems perform significantly worse then the English baseline, our gold standard. As for comparison among the cross-lingual recognizers themselves, French performed best while German performed worst. Apparently the phoneme coverage shown in table 1 does not seem to have a strong influence on the quality of the segmentation. This is also reflected in a Pearson coefficient of 0.562 for the correlation between a language's coverage and the respective monolingual system's F1 score. We can also see in table 2 that while individual systems may perform slightly higher than the multilingual one, the latter does outperform the majority of them (see discussion in section 5.6).

5.3. Results on clean English speech

We also applied our systems to audio taken from the TIMIT corpus. The speech samples provided by TIMIT differ in two major ways from the English audio in Euronews. Firstly, TIMIT consists of individual sentences recorded in a studio environment specifically for the purpose of building the corpus. The audio is therefore free of any background noise. Secondly, TIMIT features speakers of American English, whereas Euronews broadcasts feature British English.

Results indicate that the system does not seem to generalize sufficiently to be used on this data. As can be seen in Figure 2, the F-Score steadily decreases as the number of boundaries predicted by the system decreases. This is the case even for the phoneme recognizers trained on the English Euronews audio. The behaviour of the F1 curves seems to stem from a quasi-random precision in predicting phoneme boundaries. Figure 3 shows the precision curves of three recognizers, compared to a baseline representing the expected precision when guessing borders randomly. Since the behaviour of the precision curve is near-flat, the F1 curve is dominated by recall, which naturally converges towards 1 with an increasing number of generated boundaries.

There seems to be an acoustic mismatch here which indicates that the recording environment for similar experiments, as well as practical application, must be carefully chosen such that source and target data constitute a good fit.

5.4. Results on English speech using language models

After these initial experiments, we trained a phoneme level language model using the annotations provided by the Euronews corpus for our five source languages. For this purpose we first converted the orthographic transcriptions using the G2P component of MARY [15], then estimated a language model on the resulting phoneme strings.

Substituting the uniform LMs for these new models with-

out further adjustments resulted in considerably slower decodings on the English Euronews audio. For the multilingual recognizer this meant decoding at a factor of approximately 200 times real time. This is most likely caused by the significantly larger search space introduced by the length of the phoneme sequences per utterance. Unfortunately there was very little improvement in performance gained in return, in both cross- and multilingual application. It should be noted that due to the above-mentioned decrease in speed, experiments using estimated language models were only run on a subset of the training data used in 5.2, featuring approximately 17% less audio, so that scores are not directly comparable. We therefore provide scores derived from experiments on the same subset in Table 3.

System	without LM	with LM
EN (baseline)	0.7769	0.7484
M5	**0.6624**	**0.6708**
DE	0.6130	0.6446
M5*	n/a	0.5908

Table 3: Results with and without language models

When reducing the search space by narrowing the search beams, decoding became significantly faster, but this also noticeably impacted the performance indicated by the F-score (indicated as M5* in Table 3). Therefore we must assume that the manner in which the phoneme language models were trained is not suitable for crosslingual application.

5.5. Results on Basaa speech

Finally we ran our experiments on the Basaa language data described in section 4.1. The results are displayed in table 4. Unlike with our experiments on English, there is no Basaa language recognizer the authors of this paper are aware of that could serve as a baseline. However, absolute performance, as indicated by F-Scores, is lower than that on English audio, as seen in Table 2.

It should be noted that the Basaa audio used here is also free of noise, just as the TIMIT audio. Unlike with TIMIT though, performance as indicated by Precision, Recall and F-Score is not random but behaves largely as expected across different SP values. This suggests that the behaviour on TIMIT data can't originate from the absence of noise in those recordings alone (see section 5.3).

System	Precision	Recall	F-Score
M5	**0.4730**	**0.5385**	**0.5036**
DE	0.4658	0.5170	0.4900
FR	0.5166	0.5149	0.5158
IT	0.4808	0.5209	0.5000
RU	0.4710	0.6504	0.5463
TR	0.4891	0.5538	0.5195

Table 4: Comparison of segmentation quality on Basaa audio

5.6. Discussion

The results presented in this paper are encouraging regarding the use of multilingual recognizers for the given task of phoneme segmentation on a previously unseen language. The multilingual recognizer in all cases performed about as well as any of the monolingual systems. While there are monolingual systems that perform better on English *or* Basaa audio, the performance is not consistent across target languages. For example, while Russian performed best on Basaa, it also showed the second-worst performance on English. Since in practical application there is no way of predicting which source language might perform best individually, using a multilingual system instead will very likely lead to a more consistent expected result. Further experiments with more source/target pairs could confirm this assumption with higher certainty.

6. Conclusion

In this paper we have evaluated the cross- and multilingual use of phoneme recognition systems for phoneme segmentation of previously unseen languages. To this end we trained regular mono- and multilingual recognizers on noisy television news audio. We then evaluated the segmentations produced by these systems on English and Basaa audio, pretending we had no information about English as a *faux* unseen target language. We have presented our results, which show that the monolingual recognizer was able to predict segmentation boundaries with some reliability. Our results also indicate that the additional information gained from adding languages to the training data for the acoustic model of a multilingual recognizer can positively influence performance, while at least making it more robust across multiple target languages. In any case, while a multilingual acoustic model did contribute positively in this manner, adding a multilingual phoneme language model estimated on training data taken from the same source languages did not noticeably do so.

In our experiments we used noisy audio with pervasive background noise accompanying the actual speech as training data. In a real-life scenario of exploring and documenting previously unknown languages it stands to reason that recordings of target audio would be performed in a controlled environment with negligible noise. As we have shown the systems we trained did not generalize well in some cases of clean target audio (TIMIT), but did so in others (Basaa). Therefore further experiments are required to investigate if the same approach using clean training audio could potentially yield better results. Alternatively, additional filtering steps during training may improve performance without restricting the process to using acoustically clean data.

As for the influence of phoneme coverage on performance, a failure analysis could determine if there is a correlation between incorrectly predicted boundaries and whether the specific phoneme to which they belong was covered by the source language(s) or not.

While it cannot be stated with certainty yet whether the performance of the chosen approach is sufficient for the intended use case of documenting unseen languages, the results presented here can serve as a baseline for future experiments and refinement. For example, one could compare performance with a system that uses voting on boundaries among multiple monolingual decoders instead of mixed multilingual models.

7. Acknowledgements

This work was realized in the framework of the ANR-DFG project BULB (ANR-14-CE35-002).

8. References

[1] R. G. G. Jr. and B. F. G. (Eds.), *Ethnologue: Languages of the World*. Dallas, Texas, USA: SIL International, 2005.

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

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

[4] S. S. Gilles 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, A. Rialland, M. V. de Velde, F. Yvon, and S. Zerbian, "Breaking the Unwritten Language Barrier: The BULB project," in *Proceedings of the 5th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'16)*, 2016.

[5] O. Scharenborg, V. Wan, and M. 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.

[6] C. ying Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 40–49.

[7] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 3989–3992.

[8] J. wei Kuo, H. yi Lo, and H. min Wang, "Improved HMM/SVM methods for automatic phoneme segmentation," in *Proceedings of Interspeech*, 2007, pp. 2057–2060.

[9] M. Versteegh, 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.

[10] P. K. Muthukumar and A. W. Black, "Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2613–2617.

[11] P. Baljekar, S. Sitaram, P. K. Muthukumar, and A. W. Black, "Using articulatory features and inferred phonological segments in zero resource speech processing," in *16th Annual Conference of the International Speech Communication Association*, 2015.

[12] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, August 2001.

[13] M. Woszczyna, N. Aoki-Waibel, F. D. Bu, 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 *Proceedings of the 1994 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Adelaide, Australia, 1994.

[14] H. Soltau, 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*. IEEE, 2001, pp. 214–217.

[15] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[16] P. M. Lewis, G. F. Simons, and C. D. Fennig, Eds., *Ethnologue: Languages of the world*, 18th ed. Dallas, Texas: SIL International, 2015.

[17] R. Gretter, "Euronews: a multilingual benchmark for ASR and LID," in *15th Annual Conference of the International Speech Communication Association*, 2014.

[18] J. Garofolo *et al.*, "Timit acoustic-phonetic continuous speech corpus ldc93s1," Web download, Linguistic Data Consortium, Philadelphia, USA, 1993.