CROSS-LINGUAL LEXICAL LANGUAGE DISCOVERY FROM AUDIO DATA USING MULTIPLE TRANSLATIONS

F. Stahlberg[1,2], *T. Schlippe*[1], *S. Vogel*[2], *T. Schultz*[1]

[1] Cognitive Systems Lab, Karlsruhe Institute of Technology, Germany
[2] Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar

ABSTRACT

Zero-resource Automatic Speech Recognition (ZR ASR) addresses target languages without given pronunciation dictionary, transcribed speech, and language model. *Lexical discovery* for ZR ASR aims to extract word-like chunks from speech. Lexical discovery benefits from the availability of written translations in another source language [1, 2, 3]. In this paper, we improve lexical discovery even more by combining multiple source languages. We present a novel method for combining noisy word segmentations resulting in up to 11.2% relative F-score gain. When we extract word pronunciations from the combined segmentations to bootstrap an ASR system, we improve accuracy by 9.1% relative compared to the best system with only one translation, and by 50.1% compared to monolingual lexical discovery.

***Index Terms*—** Lexical language discovery, zero-resource automatic speech recognition, word-to-phoneme alignment, non-written languages

1. INTRODUCTION

In recent years, ASR research has shifted its focus to under-resourced conditions [4, 5, 6, 7, 8] to address less prevalent languages. Zero-resource (ZR) ASR goes one step further and even refrains from assuming the availability of a pronunciation dictionary, transcribed audio data, and texts to estimate language models in the target language [9] – only untranscribed audio data are available in the target language. Language discovery for ZR ASR consists of two steps: 1. *Phonetic discovery* finds subword units suitable for acoustic modeling. 2. *Lexical discovery* identifies word-like structures and phrases based on phonetic transcriptions of continuous target language speech. *Word segmentation* from audio refers to segmenting the complete phonetic target language transcriptions into word-like units and thus is a form of lexical discovery.

Our research is closely related to lexical discovery: We aim to identify words and their pronunciations in the target language to use them for ASR. Similarly to ZR ASR, no transcriptions, text resources and pronunciation dictionary for the target language are available. However, in contrast to ZR ASR, we assume the availability of an additional knowledge source: We have access to written translations in a source language of the untranscribed target language audio data. By using them we find significantly better word segmentations [1, 2, 3] compared to the monolingual case [9, 10, 11, 12]. Previous work has shown that cross-lingual lexical discovery with only one translation is applicable to a variety of corpora and languages (e.g. BTEC [1, 2, 13], BMED [14], and Christian Bible [3, 13]). In this paper, we present an algorithm which uses the synergies of multiple translations and improves word segmentation and ASR performance in the target language.

If the translations are not already at hand (e.g. in multilingual parliaments), they can be produced by a human translator and transcribed by an ASR system for the resource-rich source language. If only one translation exists in a source language, further translations can be generated automatically since Machine Translation (MT) systems exist for several languages today. Furthermore, we propose to leverage massively parallel texts (MPTs), i.e. parallel texts that are available in more than two languages [15]. In addition to the Christian Bible [16] (translated in over 513 languages [17]), many other MPTs are publicly available [15, 18, 19, 20]. [21, 22] present methods to utilize the high parallelism in MPTs for speech processing but not in the context of language discovery.

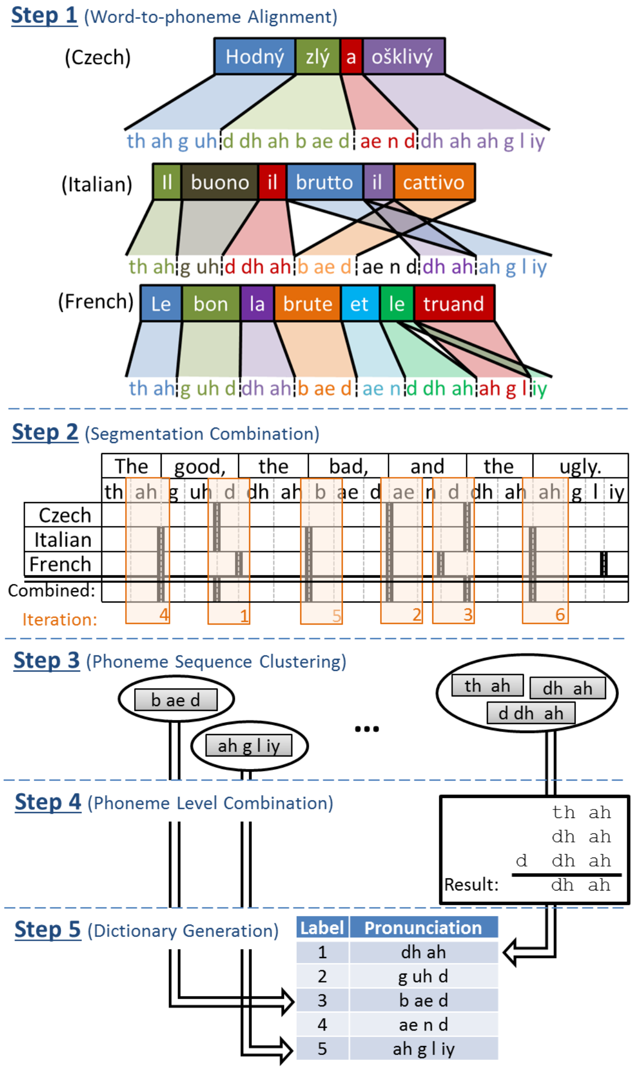
2. LEXICAL LANGUAGE DISCOVERY USING MULTIPLE TRANSLATIONS

Fig. 1 shows our complete method for discovering words and their pronunciations using multiple translations. The target language audio is represented as phoneme sequence (Sec. 3.1). The steps are described in the following sections.

2.1. Cross-lingual word-to-phoneme alignments

Our first step is to obtain multiple word segmentations using each available source translation separately. The word segmentation can be derived from an alignment between words in the source language and phonemes in the target language. Such alignments can be found with the word-to-word aligner GIZA++ [23, 1]. However, our PISA Alignment Tool[1] [2]

[1] Available at `http://pisa.googlecode.com/`


Step 1 (Word-to-phoneme Alignment)
(Czech)
Hodný zlý a ošklivý
th ah g uh d dh ah b ae d ae n d dh ah ah g l iy
(Italian)
Il buono il brutto il cattivo
th ah g uh d dh ah b ae d ae n d dh ah ah g l iy
(French)
Le bon la brute et le truand
th ah g uh d dh ah b ae d ae n d dh ah ah g l iy
Step 2 (Segmentation Combination)
The good, the bad, and the ugly.
th ah g uh d dh ah b ae d ae n d dh ah ah g l iy
Czech
Italian
French
Combined:
Iteration:
4
1
5
2
3
6
Step 3 (Phoneme Sequence Clustering)
b ae d
ah g l iy
...
th ah
dh ah
d dh ah
Step 4 (Phoneme Level Combination)
th ah
dh ah
d dh ah
Result:
dh ah
Step 5 (Dictionary Generation)

Label	Pronunciation
1	dh ah
2	g uh d
3	b ae d
4	ae n d
5	ah g l iy

Fig. 1. System overview.

is more suitable for word-to-phoneme alignments since the implemented alignment model has been particularly designed for phoneme sequences on the target side. [13] contains in-depth discussions of PISA on various corpora and languages.

2.2. Word segmentation combination

Word segmentations from multiple source translations can be combined by position-wise voting: At each position, we insert a word boundary if the number of source segmentations detecting a word boundary is above ϵ (*required votes*). Tab. 1 shows two possible issues: If ϵ is below 40%, two boundaries are inserted into the final segmentation, despite that all source segmentations contain only one. If ϵ is over 60%, no boundary is inserted although all segmentations agree that a word boundary is roughly at this position. Therefore, we introduce a position tolerance to the voting. A tolerance of 1 worked best in our experiments. We add word boundaries to the final segmentation gradually by repeating the following steps:

a) Scan the utterance from left to right. Find the first-best window position *pos* which maximizes the number of source segmentations with a word boundary inside the window. To realize a tolerance of 1, we use a window width of 2 (i.e. the window covers two possible word boundary positions).

b) If the number of source segmentations with boundaries at *pos* or $pos + 1$ is below ϵ (*required votes*), output current segmentation of the whole utterance and terminate.

c) Otherwise: To choose between the two positions inside the window (*pos* and $pos+1$), add the position with a higher number of boundaries in the source segmentations to the final segmentation.

d) Delete all boundaries already involved in the voting in *Step c)* from the source segmentations so that the algorithm finds another position in the next iteration. Return to *Step a)*.

Step 2 in Fig. 1 illustrates this algorithm. The windows are highlighted in orange. Windows 1-3 (first three iterations) cover word boundaries in all source segmentations (*Czech*, *Italian*, *French*). In the 4th iteration, all boundaries in the windows 1-3 have been deleted in *Step c)*. Consequently, a boundary is inserted where two of three segmentations agree – i.e. at window 4 (all but *Czech*). After the 6th iteration, only one single sparse boundary remains in the source segmentation *French* causing the algorithm to terminate since $\epsilon = \frac{2}{3}$ in our example (i.e. two source segmentations need to agree).

2.3. Phoneme sequence clustering

Instead of using the phoneme sequence segments in the word segmentation directly, a clustering is usually applied to compensate for phoneme recognition and alignment errors [3, 14] as illustrated in step 3 in Fig. 1. We apply the k-means algorithm and set $k = 12,000$ as this is the vocabulary size of our evaluation set EN_{filt} (introduced in Sec. 3.1). In reality the target language vocabulary size is unknown but may be estimated with the vocabulary sizes of the source languages [14].

2.4. Phoneme level combination

We extract a single word pronunciation for each cluster (step 4 in Fig. 1). For a more detailed description of this phoneme-level combination approach, we refer to [3].

Source Segmentation *es3*:	w o r d s │ e g m e n t a t i o n
Source Segmentation *es2*:	w o r d │ s e g m e n t a t i o n
Source Segmentation *pt2*:	w o r d s │ e g m e n t a t i o n
Source Segmentation *fr2*:	w o r d │ s e g m e n t a t i o n
Source Segmentation *de1*:	w o r d │ s e g m e n t a t i o n
Position-wise Voting, $\epsilon \leq 40\%$:	w o r d │ s │ e g m e n t a t i o n
Position-wise Voting, $\epsilon > 60\%$:	w o r d s e g m e n t a t i o n
Position tolerance of 1 (our method):	w o r d │ s e g m e n t a t i o n

Table 1. Position-wise voting with off-by-one errors.

ID	Source Language	Full Source Language Bible Version Name	F-score (in %)
es3	Spanish	La Biblia de las Américas	77.5
es2	Spanish	Reina-Valera 1960	74.2
pt2	Portuguese	João Ferreira de Almeida Atualizada	73.2
fr2	French	Louis Segond	72.9
de1	German	Schlachter 2000	72.1
de2	German	Luther Bibel	72.0
it	Italian	Nuova Riveduta 2006	71.8
fr1	French	Segond 21	67.6
da	Danish	Dette er Biblen på dansk	67.4
pt1	Portuguese	Nova Versão Internacional	66.7
es1	Spanish	Nueva Versión Internacional	63.5
bg	Bulgarian	Bulgarian Bible	64.1
se	Swedish	Levande Bibeln	51.7
cs	Czech	Bible 21	51.6

Table 2. Overview of the used Bible translations.

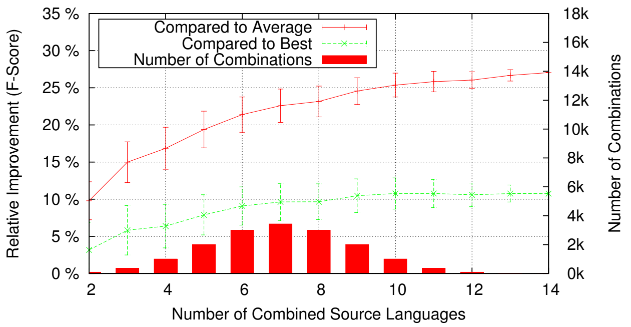


Fig. 2. Improvements on error-free phoneme sequences with $\epsilon = 25\%$. The error bars show the standard derivation.

3. EXPERIMENTS

3.1. Experimental setup

We evaluated our methods on the Christian Bible. English (English Standard Version (ESV) [24] Bible) took the role of the under-resourced target language to give us a deeper insight in the strengths and weaknesses of our algorithms. We extracted a corpus with 30.6k parallel verses. We refer to the English portion as EN_{all}. We describe this corpus in [3].

In initial experiments, we replaced the words in the target language with their canonical pronunciations. This results in error-free phoneme sequences (0% phoneme error rate (PER)) with word boundary markers which serve as our ground-truths. The pronunciations were taken from the CMUdict [25] (39 phonemes) or generated with a grapheme-to-phoneme model trained on it. Tab. 2 contains all source translations and the F-score when they are used to segment error-free English phoneme sequences.

GIZA++ first calculates initial alignments which are then refined by PISA [2]. Due to restrictions of GIZA++, we extracted the subcorpus EN_{filt} with 23k verses shorter than 101 phonemes and a ratio between source language words and target language phonemes lower than 1:12. For English[2], Crossway provides recordings of a single male speaker[3]. All recordings (EN_{all}) have a total length of 67:16h. The subcorpus EN_{filt} contains 40:28h speech. We trained a context- and speaker-dependent acoustic model (AM) on EN_{all} without EN_{filt} ($EN_{all} \setminus EN_{filt}$) with 39 phonemes (26:48h). Our phoneme recognizer using this AM and a phoneme-level 3-gram LM achieves 13.1% PER on EN_{filt}. Although the AM represents an oracle experiment since it is trained on target language transcriptions, we feel that it is sufficient for a proof of concept for our methods. Cross-lingual word-to-phoneme alignment and pronunciation extraction with significantly higher PERs have been studied in [13, 14]. For unsupervised acoustic modeling and phonetic language discovery, we refer to [26, 27, 28] and related papers.

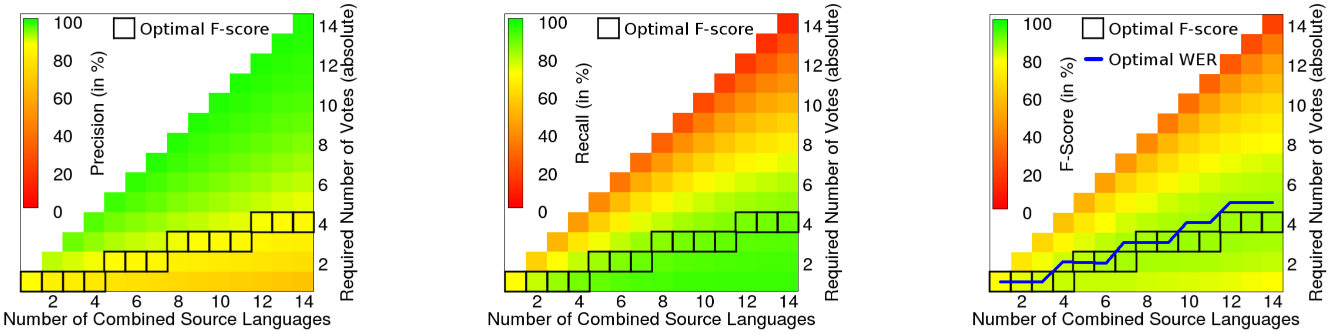
3.2. Word segmentation

Fig. 2 plots the relative improvements compared to both the best and the average F-score of the combined source translations when the translations are selected arbitrarily. The red bars visualize the number of possible source language combinations – e.g. there are $\binom{14}{10} = 1,001$ distinct 10-combinations of 14 source languages. The results are generated with error-free phoneme sequences since no ground-truth for recognized phoneme sequences are available. The green curve saturates at about 10% mean relative improvement. The standard derivation decreases with the number of translations.

In our further analysis, we combined the source translations with the best F-score instead of all possible subsets for simplification. Fig. 3 visualizes the impact of the ϵ-parameter (Sec. 2.2) to precision, recall, and F-score on error-free phoneme sequences. The vertical axis is related to the ϵ-parameter but represents the absolute required number of votes (in contrast to ϵ denoting the required vote share relative to the total number of source languages). Black borders highlight the cells corresponding to the optimal F-score. Yellow represents the best *mono*lingual segmentation (*es3*), green improvements and red declines compared to it. With a high ϵ, the word boundaries are likely to be correct (high precision) but many boundaries are missing (low recall, tendency for under-segmentation). However, a low ϵ leads to more word boundaries (high recall, over-segmentation), whereas a larger fraction of them is incorrect (low precision). Even though $\epsilon = 25\%$ optimizes the F-score, it only significantly improves the recall – The highlighted squares in Fig. 3(a) are sometimes reddish indicating a decline in precision. If we increase the absolute number of required votes by 1 (squares just above the marked squares in Fig. 3), both precision and recall are usually improved but the F-score is not optimal.

[2] For audio recordings in a variety of other languages see e.g. `http://www.bible.is/` or `http://www.talkingbibles.org/`

[3] Available for purchase at `http://www.crossway.org/bibles/esv-hear-the-word-audio-bible-610-dl/`



(a) Precision ($\frac{\text{true positive}}{\text{true positive+false positive}}$) (b) Recall ($\frac{\text{true positive}}{\text{true positive+false negative}}$) (c) F-score ($2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision+recall}}$)

Fig. 3. Segmentation quality for the combination of multiple source segmentations.

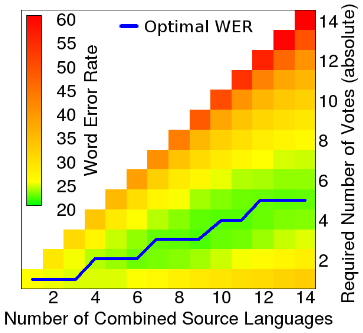


Fig. 4. ASR performance for segmentation combination.

3.3. Automatic speech recognition

It turns out that improving both precision and recall is more important than improving the F-score when the dictionary is used in an ASR system. We replaced the segments in the segmented phoneme sequences with the closest entry in the extracted dictionary to train a unigram LM as described in [14]. Using higher order n-gram LMs did not improve results. The green area in Fig. 4 demonstrates that our segmentation combination effectively improves ASR performance if we set ϵ correctly. The ascending slope of the green band indicates that $\epsilon \approx 33\%$ is optimal for the WER (blue line in Fig. 3(c) and 4). WER reductions (green areas in Fig. 4) correspond to simultaneous improvements in precision and recall (intersection of green areas in Fig. 3(a) and (b)).

Fig. 5 illustrates the ASR improvements through our segmentation combination. The curves are rather flat for a low number of combined source translations. They approach a U-shape as the number of translations increases. The minimum of the curves improves with increasing number of translations, i.e. adding more translations helps to improve the WER. However, the improvements saturate with more than eight translations: The minima for 8 (yellow), 11 (black), and 14 (orange) translations are approximately at the same level.

Tab. 3 summarizes the WERs with recognized phoneme sequences. Without source language translation, we achieve 59.9% WER with a monolingual word segmentation method like Adaptor Grammars [10]. When the *es3* translation is available, we can reduce the WER to 32.9% using PISA. The combination of nine source translations leads to 29.9% WER.

4. CONCLUSION

We studied lexical language discovery for ASR from audio with the help of multiple written translations. If the translations are not at hand, they can be produced by a human translator and generated automatically for further languages using MT. Our algorithm for word segmentation combination achieves up to 11.2% relative gain in F-score. For ASR we get 9.1% relative improvement in WER compared to the best system with only one translation, and 50.1% compared to monolingual lexical discovery on the Christian Bible. Our method has the potential to tackle non-written languages for ASR since we do not require a target language writing system.

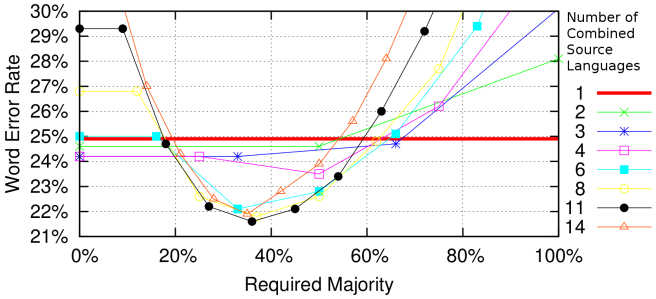


Fig. 5. WER for segmentation combination over ϵ.

Required Resource	Method	WER (in %)
target language phoneme sequence	Adaptor Grammars (*colloc2* grammar)	59.9
+ 1 source translation translation (*es3*)	GIZA++	37.3
	PISA	32.9
+ 8 additional translations	segmentation combination ($\epsilon = 33\%$)	29.9

Table 3. WER with different methods and resources.

5. REFERENCES

[1] S. Stüker and A. Waibel, "Towards Human Translations Guided Language Discovery for ASR Systems," in *SLTU*, 2008.

[2] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word Segmentation through Cross-Lingual Word-to-Phoneme Alignment," in *SLT*, 2012.

[3] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Pronunciation Extraction from Phoneme Sequences through Cross-Lingual Word-to-Phoneme Alignment," in *SLSP*, 2013.

[4] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic Speech Recognition for Under-Resourced Languages: A Survey," *Speech Communication*, 2014.

[5] T. Schlippe, S. Ochs, and T. Schultz, "Web-Based Tools and Methods for Rapid Pronunciation Dictionary Creation," *Speech Communication*, 2014.

[6] H. Gelas, S. T. Abate, L. Besacier, F. Pellegrino, et al., "Quality Assessment of Crowdsourcing Transcriptions for African Languages," in *Interspeech*, 2011.

[7] S. Kanthak and H. Ney, "Context-Dependent Acoustic Modeling Using Graphemes for Large Vocabulary Speech Recognition," in *ICASSP*, 2002.

[8] T. Schultz and A. Waibel, "Language-independent and Language-adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, 2001.

[9] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, et al., "A Summary of the 2012 JHU CLSP Workshop on Zero Resource Speech Technologies and Models of Early Language Acquisition," in *ICASSP*, 2013.

[10] M. Johnson and S. Goldwater, "Improving Non-Parameteric Bayesian Inference: Experiments on Unsupervised Word Segmentation with Adaptor Grammars," in *HLT-NAACL*, 2009.

[11] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, "Unsupervised Word Segmentation from Noisy Input," in *ASRU*. IEEE, 2013.

[12] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling," in *ACL*, 2009.

[13] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word Segmentation and Pronunciation Extraction from Phoneme Sequences Through Cross-Lingual Word-to-Phoneme Alignment," *Computer Speech & Language*, 2014.

[14] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Towards Automatic Speech Recognition Without Pronunciation Dictionary, Transcribed Speech and Text Resources in the Target Language Using Cross-Lingual Word-to-Phoneme Alignment," in *SLTU*, 2014.

[15] M. Cysouw and B. Wälchli, "Parallel Texts: Using Translational Equivalents in Linguistic Typology," *STUF-Sprachtypologie und Universalienforschung*, 2007.

[16] T. Mayer and M. Cysouw, "Creating a Massively Parallel Bible Corpus," *Oceania*, 2013.

[17] Bob Creson, "Wycliffe Bible Translators," `http://wycliffe.org/`, 2014, Accessed on 4th October 2014.

[18] A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel, "The AMARA Corpus: Building Parallel Language Resources for the Educational Domain," in *LREC*, 2014.

[19] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *MT Summit*, 2005.

[20] B. Wälchli, *Co-Compounds and Natural Coordination*, Oxford University Press, 2005.

[21] T. Mayer and M. Cysouw, "Language Comparison Through Sparse Multilingual Word Alignment," in *EACL*, 2012.

[22] R. Östling, "Bayesian Word Alignment for Massively Parallel Texts," in *EACL*, 2014.

[23] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, 2003.

[24] Crossway, "The Holy Bible: English Standard Version," 2001.

[25] R Weide, "The Carnegie Mellon Pronouncing Dictionary 0.6," 2005.

[26] C. Lee and J. Glass, "A Nonparametric Bayesian Approach to Acoustic Model Discovery," in *ACL-HLT*, 2012.

[27] M. J. F. Gales, K. M. Knill, A. Ragni, and P. R. Rath, "Speech Recognition and Keyword Spotting for Low Resource Languages: Babel Project Research at CUED," in *SLTU*, 2014.

[28] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised Learning of Acoustic Sub-Word Units," in *ACL-HLT*, 2008.