TOWARD HUMAN-ASSISTED LEXICAL UNIT DISCOVERY WITHOUT TEXT RESOURCES

Chris Bartels, Wen Wang, Vikramjit Mitra, Colleen Richey, Andreas Kathol, Dimitra Vergyri
Harry Bratt, Chiachi Hung

SRI International, Menlo Park, CA, U.S.A.

ABSTRACT

This work addresses lexical unit discovery for languages without (usable) written resources. Previous work has addressed this problem using entirely unsupervised methodologies. Our approach in contrast investigates the use of linguistic and speaker knowledge which are often available even if text resources are not. We create a framework that benefits from such resources, not assuming orthographic representations and avoiding generation of word-level transcriptions. We adapt a universal phone recognizer to the target language and use it to convert audio into a searchable phone string for lexical unit discovery via fuzzy sub-string matching. Linguistic knowledge is used to constrain phone recognition output and to constrain lexical unit discovery on the phone recognizer output.

Target language speakers are used to assist a linguist in creating phonetic transcriptions for the adaptation of acoustic and language models, by respeaking more clearly a small portion of the target language audio. We also explore robust features and feature transform through deep auto-encoders for better phone recognition performance.

The proposed approach achieves lexical unit discovery performance comparable to state-of-the-art zero-resource methods. Since the system is built on phonetic recognition, discovered units are immediately interpretable. They can be used to automatically populate a pronunciation lexicon and enable iterative improvement through additional feedback from target language speakers.

Index Terms: lexical discovery, low resource languages, automatic speech recognition

1. INTRODUCTION

State-of-the-art automatic spoken language technology tasks that involve recognizing words from acoustic input, such as automatic speech recognition and spoken term discovery, typically rely on linguistic resources such as phoneme inventories, pronunciation dictionaries, and annotated speech data. Such resources are unavailable for many languages and expensive to create.

Recent work on zero-resource spoken language learning has addressed the scenario in which no resources are available for system development [1, 2, 3, 4]. That research targeted unsupervised phonetic and lexical unit discovery from audio data and produced some interesting results. Nevertheless, these approaches still result in significantly degraded performance compared to using written resources with standardized orthography.

Alternatively, other researchers have concentrated on developing systems using a small amount of transcribed data, aiming to achieve acceptable performance, while still minimizing annotation expense. In particular, the IARPA Babel program extended speech recognition capabilities to under-resourced languages, targeting the task of keyword spotting (KWS) in audio documents [5, 6]. However, these efforts still rely on significant amounts of manual audio transcriptions and lexicons with standardized orthography, which is an expensive and time-consuming activity even for a small number of words.

Real-world scenarios often lie in the space between zero-resource and limited-resource approaches. Although many languages lack written resources, they have existing linguistic studies. Typically, researchers and system developers also have access to speakers with some proficiency in the target language [7, 8], as well as to a wide variety of annotated data in other languages.

This work addresses lexical discovery in languages without written development resources by creating a framework that benefits from those limited but commonly available linguistic and speaker resources. Lexical discovery is implemented as a search of repeated phone strings, by using fuzzy substring matching on the output of a universal phone recognizer. Unlike several of the lexical discovery approaches in the literature that work with pattern matching on the acoustic space or use automatically discovered subword units, we chose to work with phone units because it enables leveraging linguistic knowledge and using input from linguists to improve system performance.

This approach achieves performance on lexical discovery that is comparable to competing zero-resource methods based on acoustic-only pattern matching. Because the system is built on phonetic recognition, the results are immediately interpretable by linguists, can be used to seed a pronunciation lexicon, and offer the potential for iterative improvement through additional feedback from the target language speaker.

Contributions of our human-assisted lexical unit discovery technique include:

- We explored linguistic knowledge and respeaking data collection with the assistance of native speakers to adapt and refine the universal phone recognizer and achieved significant improvement on the phone recognition performance. We also explored the linguistic knowledge in lexical unit discovery based on fuzzy substring matching and achieved improvement on the lexical unit discovery performance.
- We investigated the effectiveness of noise robust features for lexical unit discovery in the dynamic time warping (DTW) framework and observed a significant improvement on performance compared to the mel-frequency cepstral coefficient (MFCC) features, especially a significant improvement on the cross-speaker lexical unit discovery performance.

2. APPROACH: HUMAN-ASSISTED SYSTEM DEVELOPMENT

This section outlines our approach to human-assisted lexical discovery. Each aspect of the approach is summarized here, and described in detail in the following sections.

- A universal phone recognizer is used to decode the audio into a string of phones. This universal recognizer was created using supervised training from a variety of languages with resources.
- As expected, the universal phone recognizer's initial performance was fairly poor for unseen languages. Our next step was to improve the recognition performance using linguistic knowledge that is typically available even when text resources are not. In particular, the language's phonetic inventory and syllable structure are used to constrain the recognition output.
- We also had access to target language speakers, and we used them as a resource to improve recognition performance. We recorded the native speakers re-speaking a small amount of the real-world target language data. Compared to the real-world recordings, the respoken data was more intelligible and had less background noise, which allowed a trained linguist to manually create phonetic transcriptions. These transcriptions were then used to adapt the recognizer to the target language.
- The baseline for our lexical discovery experiments was to search for repeated patterns directly on the acoustics using dynamic time warping (DTW). We used the algorithms and publicly available tools from [9, 10]. Our results were evaluated using the metric presented in [11].
- We additionally investigated the effectiveness of noise robust features for lexical unit discovery in the DTW framework.
- Finally, our lexical discovery approach finds words by searching for repeated phone strings in the output of the universal phone recognizer. We used the standard repeated substring detection algorithm, and also explored several modifications.

3. EXPERIMENTAL SETUP

We evaluate our results on the Amharic and Pashto development sets from the IARPA Babel program. This audio was collected under a variety of real-world conditions and contains background noise. Amharic has approximately 7 hours of audio and 50,000 words, and Pashto has 8 hours and 100,000 words. Lexica and transcriptions are available for these languages, and these were used to evaluate our experimental systems that do no rely on such resources. The results of both languages were analyzed and optimized during development, so all results should be viewed as development set results.

4. UNIVERSAL PHONE RECOGNIZER

We begin the detailed descriptions of our approach with the universal phone recognizer. The acoustic and language models of the recognizer were trained using seven language corpora from different sources: Assamese (Babel), Bengali (Babel), Dari (Transtac), Egyptian Arabic (Callhome), English (Fisher), Mandarin (DARPA GALE program), and Spanish (Callhome). This gave approximately 650 hours of audio.

We scored the phonetic recognition against phonetic references generated from a forced alignment. All of our results are time-mediated phone error rates, which were typically 3%–8% absolute worse than the string-aligned phone error rates. Time-mediated scoring was used because when dealing with high error rates and a vocabulary of 58 phones, string alignment often introduced noise through spurious alignments.

The phone set we used distinguishes most of the sounds described as phonemically contrastive in the languages, but it also merges acoustically similar sounds that may be contrastive. It does not distinguish pulmonic and non-pulmonic consonants—thus, the ejective consonants of Amharic are not distinguished from their pulmonic counterparts. Consonants with a secondary place of articulation are not distinguished from consonants with only a primary place of articulation—thus, the pharyngealized consonants of Arabic are not distinguished from their plain counterparts. The phone set distinguishes among all manners of articulation, but it merges several contrasts among places of articulation. For example, it does not distinguish the place of articulation between dental, alveolar, and retroflex stops. The set contained 55 speech phones and 3 non-speech phones.

All acoustic models (AMs) were deep neural networks (DNN) optimized for cross entropy with clustered triphone targets. For reported results, we used DNN with 5 hidden layers and 1200 neurons per layer. The front-end for most results was 13 mel-frequency cepstral coefficient (MFCC) features along with deltas and double deltas, and feature-space maximum likelihood linear regression (fMLLR) was used to adapt these features to the target data (unsupervised) [12]. The results labeled "Language Specific" used 40 mel-scaled filterbanks.

The language models (LMs) were trained on phonetic transcriptions generated with forced alignments. We used bi-grams, as larger n-grams performed worse. A bi-gram was trained separately for each training language, and the seven individual bi-grams were combined with uniform interpolation. Silence phones were included in the language model (instead of uniform probability insertions as is typically done in word decoding).

5. LINGUISTIC CONSTRAINTS

As discussed, many real-world languages without written resources have existing linguistic studies. Studies typically describe the language's phonetic inventory, and our approach allows us to easily take advantage of this by constraining the output of the recognizer to the target language's phone set.

Decoding can be further constrained by taking the language's syllable structure into account. Amharic has a particularly well-defined syllable structure. It allows no more than one consonant in the syllable onset position (except labialized consonants), and no more than two consonants in other positions (e.g., (C)V(C)(C)). Pashto allows for complex onsets consisting of up to three consonants, and complex codas of up to two consonants (e.g., (C)(C)(C)V(C)(C)). The sequence of consonants in Pashto's onsets and codas is further constrained, in part, by the sonority hierarchy and other similar constraints. Even so, Pashto allows for a greater variety of syllables than Amharic.

The phonetic and syllabic constraints were defined using the Foma constraint language [13]. The constraints are compiled into finite state transducers, and the recognizer's transducer was composed with the constraint transducer to limit its output.

The results for the linguistic constraints are given in Table 1. As one might expect, we see noticeable improvement by limiting the phone set. Although this improvement is straightforward, it demonstrates how this commonly available knowledge can be utilized by our approach. The syllable constraint gives additional improvement for Amharic. The improvement for Pashto is smaller, as Pashto has more syllable variety than Amharic.

6. ADAPTATION BASED ON HUMAN INPUT

The second improvement to our recognizer came from adapting it to a small amount of hand-generated phonetic transcriptions. The

Table 1. *Phonetic recognition results for linguistic constraints. Values are time-mediated phone error rates.*

Constraint	Amharic	Pashto
None	76.3%	73.0%
Phone	75.1%	71.8%
Syllable	74.6%	71.5%

target-language audio was collected under real-world conditions, is mostly narrow bandwidth, has background noise, and has dropped samples. Creating even a small amount of phone-level transcriptions on such data is a difficult task, even for a trained phonetician. However, native speakers find the audio intelligible for the most part, and they can easily listen and respeak the utterances. We recorded native speakers respeaking the utterances with better audio quality in a quiet environment, making these recordings much easier for the linguists to transcribe phonetically. Respeaking has been shown to increase transcription accuracy in other contexts, such as language preservation [14, 15]. Three modes for repeating the utterance were used:

- *Matched* - Repeat the utterance, trying to match its speed and pronunciation.
- *Slow* - Repeat the utterance with lengthened phones and more careful pronunciation. Phonological reductions are typically suppressed.
- *Pauses* - Pronounce each word matching the speed and pronunciation of the original, but make a brief pause at each word boundary. Cross-word phonological processes are typically suppressed.

The linguist first listened to all three respoken versions and then transcribed the matched and slow versions. The slow and pauses versions were used to inform the transcription of the matched audio, but many differences remained due to phonological reduction and assimilation processes. The transcriptions for the matched and pauses audio versions were then compared to ensure that any differences were truly reflected in the pronunciation and were not due to transcription variability.

For respeaking, a set of utterances was randomly selected from the training data. A native speaker listened to each utterance and, if possible, repeated each utterance in the three modes described above. An utterance was skipped if the audio clip contained no speech, speech outside the target language, or was unintelligible. For Amharic, 34 utterances were respoken and transcribed. For Pashto, two speakers combined to respeak 92 utterances, and these were all transcribed.

The universal acoustic model was adapted by re-training the original model on the adaptation data. During adaptation, the frontend fMLLR and tied states were kept the same. The language model was adapted by using the adaptation data to train a small model, and then linearly interpolating this small model with the universal model.

The results for adapting the recognizer are given in Table 2. We first adapted the acoustic model using the original Babel audio together with the hand-generated transcriptions, and this result is labeled "Hand, Babel." By combining the Babel audio with the audio from the *matched* and *pauses* utterances recorded by our native speaker, we had 97 utterances available for Amharic adaptation and 265 for Pashto (For each transcription we used three utterances: original, matched, and pauses. Discrepancies in the totals are because a few of the respoken modes were not recorded for some utterances.) This combined set gave us additional improvement for Amharic, as shown in the rows marked "Hand, All." The results when adding language model adaptation are given by the "LM Data" column. "Hand" means adaptation with the hand-generated transcriptions.

To show the potential effect of a larger transcription effort, Table 2 also gives results for adaptation using 30 minutes of audio with forced alignment transcriptions, labeled "Forced, 30'." 30 minutes is 450 utterances for Amharic and 385 for Pashto. A language-specific recognizer was trained using the Babel audio data, transcripts, and pronunciation dictionary for the certain language. Then phonetic forced alignment was generated using the recognizer. As a limiting case, the table has results for models trained using only target language training data, labeled "Language Specific." Forced alignments from target language transcriptions were used for both the acoustic and language models, and there was about 42 hours of audio for Amharic and 96 for Pashto.

We additionally combined two of the adapted systems with the syllable constraints from Section 5. These results are labeled "Hand+gen.+const." and "Forced, 30'+const."

7. LEXICAL UNIT DISCOVERY EVALUATION AND BASELINE

Next, we break from the discussion on phone recognition to describe our lexical discovery evaluation metric, which was taken from [11]. This metric provides a rapid measurement of how well a speech representation can associate examples of the same word types and discriminate different word types, as well as to assess speaker independence of the speech representation.

This evaluation metric is calculated as follows. Using the time-aligned word reference transcripts of a target language data set, we randomly sampled a set of spoken word examples with a minimum character length of 5 in the surface form, denoted $\mathcal{W} = \{w_i\}_{i=1}^{M}$. We then randomly sampled the following four sets of word pairs $(w_i, w_j) \in \mathcal{W}x\mathcal{W}, i \neq j$ from the data set [11]:

- $\mathcal{C}_1$: Same word, same speaker (SWSP)
- $\mathcal{C}_2$: Same word, different speakers (SWDP)
- $\mathcal{C}_3$: Different word, same speaker (DWSP)
- $\mathcal{C}_4$: Different word, different speakers (DWDP)

On this same target language data set, we then apply lexical unit discovery to discover repeated patterns. For each word pair (w_i, w_j) in the four sets, we assigned a discovered repeated pattern (i.e., discovered lexical unit) s_{w_i} to w_i, and a discovered lexical unit s_{w_j} to w_j, if the discovered lexical unit s_{w_i} and s_{w_j} overlap at least 50% of the duration of w_i and w_j, respectively. For each of these potential discovered units, the distance measure $Distance(w_i, w_j)$ was computed, where distance is defined by the particular lexical unit discovery approach. If multiple discovered lexical units overlap the duration of w_i or w_j, the distance was computed as the smallest distance between the discovered lexical units. If there is no overlapping discovered lexical unit to w_i or w_j, then we assigned a distance of

Table 2. *Phonetic recognition results for adaptation. Percentages are time mediated phone error rate (TPER).*

AM Adapt	LM Adapt	Amharic	Pashto
None	None	76.3%	73.0%
Hand, Babel	None	71.8%	71.4%
Hand, All	None	70.0%	71.5%
Hand, All	Hand	68.0%	70.7%
Hand, All	Hand+const.	67.2%	70.1%
Forced, 30'	Forced, 30'	62.1%	64.3%
Forced, 30'+const.	Forced, 30'+const.	62.0%	64.5%
Language Specific		53.3%	53.8%

infinity for the word pair (w_i, w_j). Then, given a threshold τ, we can compute

$$N_k(\tau) := |(w_i, w_j) \in C_k : Distance(w_i, w_j) \leq \tau| \quad (1)$$

We then computed the precision recall for SW, SWSP, and SWDP as follows:

$$P_{SW}(\tau) = \frac{N_1(\tau) + N_2(\tau)}{\sum_{k=1}^{4} N_k(\tau)} \quad (2)$$

$$R_{SW}(\tau) = \frac{N_1(\tau) + N_2(\tau)}{|C_1| + |C_2|} \quad (3)$$

$$R_{SWSP}(\tau) = \frac{N_1(\tau)}{|C_1|} \quad (4)$$

$$R_{SWDP}(\tau) = \frac{N_2(\tau)}{|C_2|} \quad (5)$$

We then sampled through a series of values of τ and computed the Precision-Recall Breakeven (PRB) point PRB_{SP} where P_{SW} and R_{SWSP} are equal, and PRB_{DP} where P_{SW} and R_{SWDP} are equal. Note that a high PRB_{SP} value indicates a good speaker-dependent speech representation for spoken term discovery, and a high PRB_{DP} value indicates a good speaker-independent speech representation for this task.

7.1. Baseline Results

The baseline for our lexical discovery experiments was a DTW-based search for repeated acoustic patterns using the algorithms presented in [9]. Different from earlier approaches of exhaustive DTW searches across the entire similarity matrix, [9] used randomized algorithms operating on acoustic features to produce sparse approximate similarity matrices much more efficiently in memory and computational complexity. All DTW experiments were performed using the publicly available *ZRTools:zero-resource speech discovery, search, and evaluation toolkit* [10].

The distance measure, $Distance(w_i, w_j)$, is the DTW distance computed using cosine distance (one minus cosine similarity). This distance measure is used to compute PRB_{SP} and PRB_{DP} as described above. The baseline result using MFCC features is given in Table 3.

Table 4 shows the word counts and sizes of the four sets $C_i, i = 1, 2, 3, 4$ sampled from the time-aligned word transcripts of the two evaluation data sets.

7.2. Robust Features

Although MFCC features are typically the standard for speech applications, they are often not the best choice in noisy conditions. Our target language data from BABEL is collected in real-world environments and contains noise, and MFCCs are not necessarily a comparable baseline to our or any method that is more robust to noise. We address this by performing lexical discovery using DTW search using features that we have found to improve noise robustness in other applications.

The first such feature is created using a deep neural network auto-encoder bottle-neck model (DAE-BN) [16]. The DAE-BN system is a five-hidden-layer, fully connected DNN system, with the third hidden layer containing a bottleneck of sixty neurons. The remaining hidden layers had 1024 neurons. The input to the DAE-BN system is 40 gammatone filterbank energies (GFBs) with an eleven frame window, and the output is the same 40 GFBs but with a three frame window. The network is trained using speech data from a variety of languages. Features are generated from the fully trained model using the activations of the bottleneck (third) layer.

The second set of robust features are Normalized Amplitude Modulation Coeffients (NMC) [17]. The NMC feature captures and uses the amplitude modulation (AM) information from bandlimited speech signals. NMC is motivated by AMs of subband speech signals playing an important role in human speech perception and recognition. NMCs are obtained by using the approach outlined in [17], in which the features are generated from tracking the AM trajectories of subband speech signals in a time domain.

Table 3 gives lexical discovery results for DTW-based search using the BN-DAE and NMC features. We observed a significant improvement on the PRB_{SP} and PRB_{DP} scores from our robust NMC features and BN-DAE features, compared to the MFCC features. This improvement is markedly large on cross-speaker performance.

8. PHONETIC-BASED LEXICAL UNIT DISCOVERY

We now describe our approach of automatic lexical unit discovery. The approach takes the recognized phone sequence as input and discovers lexical units by finding repeated substrings. By recursively applying longest repeated substring detection algorithm using suffix array, we could efficiently detect repeated substrings. For detecting longest repeated substring using suffix array, given an input string, we first sort all of its suffixes in the suffix array. Then for each two adjacent suffixes in the sorted suffix array, we count the length of the common prefix while avoiding counting the overlapped part. The longest repeated prefix appearing first on the original input string is the longest repeated substring.

We extended the standard repeated substring detection algorithm with two modifications. First, we used the silence labels in the phone recognition output and the sentence boundaries of the audio segments to be recognized as boundaries of lexical units. Second, the

Table 3. Evaluation of MFCC features and our robust NMC features and features created using a deep neural network auto-encoder bottleneck model (DAE-BN), for lexical unit discovery.

Feature	Amharic		Pashto	
	PRB_{SP}	PRB_{DP}	PRB_{SP}	PRB_{DP}
MFCC	0.33	0.14	0.35	0.16
NMC	0.47	0.40	0.48	0.40
DAE-BN	0.42	0.38	0.46	0.39

Table 4. Statistics of the evaluation data for lexical unit discovery.

Data Set	Word Count	$\|C_1\|$	$\|C_2\|$	$\|C_3\|$	$\|C_4\|$
Amharic	47,292	3,679	123,763	386,150	1,016,263
Pashto	94,721	3,567	100,939	308,632	1,037,561

algorithm was extended to fuzzy matching by using the substitutable, deletable, and insertable phones for the target language, as defined by linguistic knowledge. We used this extension to model pronunciation variability.

We also applied linguistic knowledge of syllable phonotactics, i.e., syllable constraints, to lexical unit discovery. This utilized the same finite state transducers that were used to constrain the phone recognition output. The syllable constraints were used to filter the discovered units from fuzzy substring matching, that is, only patterns compatible with the target language phonotactics are accepted as possible "words". These syllable constraints could also help further constrain the boundaries of lexical units, in addition to the above-mentioned boundary constraints.

We described the lexical unit discovery evaluation metric in Section 7. For our phonetic-based lexical unit discovery, we also first ran our phonetic-based lexical unit discovery on the evaluation data sets, then assigned discovered repeated phone sequences to evaluation word pairs with the described procedure in Section 7. The distance measure $Distance(w_i, w_j)$ is defined as the Levenshtein distance between the two assigned discovered phone sequences of lexical units s_{w_i} and s_{w_j}, divided by the maximum of the lengths of the two phone strings hence converted to a value in [0,1].

Related work to our approach of identifying lexical units from unsegmented strings of symbols includes [18, 19, 20, 21]. Some prior work models word segmentation based on the distribution of phoneme sequences in the input (e.g., [20]). Some prior work explores synergistic interactions between multiple levels of linguistic structures (e.g., [19]).

8.1. Experiments

Table 5 shows the efficacy of employing boundary constraints from sentence boundaries and silence labels generated by the phone recognizers, exploring linguistically defined confusable phones, and applying the syllable constraints. As shown in the table, using the boundary constraints and the syllable constraints made a significant improvement on the lexical unit discovery performance, whereas using the linguistically defined confusable phones for fuzzy matching hasn't been able to make a significant impact. If not specified otherwise, the lexical unit discovery evaluation results in this paper are always based on applying all the knowledge in discovery, that is, boundary constraints, linguistically defined confusable phones, and syllable constraints.

Table 6 shows the PRB_{SP} and PRB_{DP} values for the Amharic and Pashto development sets from the un-adapted phone recognizer, the best performing phone recognizers adapted on the manually created phonetic transcriptions and employing syllable constraints, and the phone recognizers trained on the training set of the target languages. The time-mediated phone recognition error rate (TPER) for each output is also shown with PRB_{SP} and PRB_{DP}. First, we observed that our lexical unit discovery approach based on phone recognition output shows strong speaker independence, as PRB_{DP} is the same as PRB_{SP}, for both target languages. Second, we observed that PRB_{SP} and PRB_{DP} scores are correlated with the phone recognition accuracy. Third, the discovered lexical units are associated with the hypothesized phone sequences, so they are interpretable by linguists and could be explored for iterative improvement of both the recognizer and lexical discovery.

9. DISCUSSION

Our phonetic approach for lexical discovery performs similar to zero-resource methods that use acoustic-only pattern matching. Ideally, our human-assisted discovery methods would outperform existing zero-resource methods. However, unlike DTW based methods, our approach creates a pronunciation lexicon that is immediately interpretable by linguists. We have also shown that the phonetic approach is viable, and this creates an alternative path for future research and possible system combination with acoustic-only results.

To further improve our lexical unit discovery approach, we plan to (1) explore phone confusion matrices from the phone recognizer for enhancing the fuzzy substring matching; (2) explore linguistic knowledge for word and syllable composition for lexical unit discovery; and (3) importantly, explore approaches to model variability (i.e., different phone sequences for the same words) (e.g., the noisy-channel model as in [3]), and approaches to jointly model sub-lexical and lexical units (e.g., the adaptor grammar as in [19]).

We also will gather feedback from the native speakers by giving them simple tasks that evaluate the accuracy of our discovered units. We will then explore using this data to iteratively improve both the recognizer and lexical discovery.

10. ACKNOWLEDGMENTS

We used the following Babel data releases: Amharic, IARPA-babel307b-v1.0b; Assamese, IARPA-babel102b-v0.5a; Bengali,

Table 5. Efficacy of employing boundary constraints, exploring linguistically defined confusable phones, and the syllable constraints for lexical unit discovery performance.

Phone Recognizer	**Amharic**			**Pashto**		
	TPER (%)	PRB_{SP}	PRB_{DP}	**TPER (%)**	PRB_{SP}	PRB_{DP}
Baseline (NO adaptation)	77.0	0.19	0.19	74.0	0.18	0.18
+ sent boundary + silence info	77.0	0.23	0.23	74.0	0.22	0.22
++ ling. phone confusions	77.0	0.25	0.25	74.0	0.23	0.23
+++ syllable constraints	77.0	0.28	0.28	74.0	0.25	0.25

Table 6. Evaluation of lexical unit discovery performance based on un-adapted phone recognizers, best adapted phone recognizers with syllable constraints, and language-specific phone recognizers.

Phone Recognizer	**Amharic**			**Pashto**		
	TPER (%)	PRB_{SP}	PRB_{DP}	**TPER (%)**	PRB_{SP}	PRB_{DP}
Baseline (NO adaptation)	77.0	0.28	0.28	74.0	0.25	0.25
Best Adapted + const.	67.2	0.40	0.40	70.7	0.37	0.37
Language Specific	53.3	0.47	0.47	53.8	0.46	0.46

IARPA-babel103b-v0.4b; and Pashto, IARPA-babel104b-v0.4bY. FullLP training sets were used.

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government

11. REFERENCES

[1] Aren Jansen, Ken Church, and Hynek Hermansky, "Towards spoken term discovery at scale with zero resources," in *Proceedings of Interspeech*, 2010.

[2] Herman Kamper, Aren Jansen, Simon King, and Sharon Goldwater, "Unsupervised lexical clustering of speech segments using fixed dimensional acoustic embeddings," in *Proceedings of SLT*, 2014.

[3] Chia-ying Lee, Timothy O'Donnell, and James Glass, "Unsupervised lexicon discovery from acoustic input," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 389–403, 2015.

[4] "The zero resource speech challenge (special sessions at interspeech 2015)," in *Interspeech*, 2015.

[5] Stavros Tsakalidis, Ruey-Chang Hsiao, Damianos Karakos, Timothy Ng, Sudhir Ranjan, Guruprasad Saikumar, Le Zhang, Long Nguyen, Richard Schwartz, and John Makhoul, "The 2013 BBN vietnamese telephone speech keyword spotting system," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7829–7833.

[6] Anton Ragni, Kate M. Knill, Shakti P. Rath, and Mark J. F. Gales, "Data augmentation for low resource languages," in *Proceedings of Interspeech*, 2014.

[7] Marelie Davel and Etienne Barnard, "The efficient generation of pronunciation dictionaries: human factors during bootstrapping," in *ICSLP*, 2004.

[8] M. Davel and E. Barnard, "Pronunciation predication with default and refine," *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[9] Aren Jansen and Benjamin Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proceedings of ASRU*, 2011.

[10] Aren Jansen, Ben Van Durme, and Greg Sell, "Zrtools: Zero-resource speech discovery, search and evaluation toolkit," https://github.com/arenjansen/ZRTools, Accessed: 2017-07-25.

[11] Michael A. Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proceedings of Interspeech*, 2011.

[12] M.J.F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[13] Hulden Mans, "Foma: a finite-state compiler and library," in *Proceedings of the EACL 2009 Demonstrations Session*, 2009, pp. 29–32.

[14] Anthony C. Woodbury, "Defining documentary linguistics," *Language Documentation and Description*, vol. 1, 2003.

[15] Steven Bird, Florian R Hanke, Oliver Adams, and Haejoong Lee, "Aikuma: A mobile app for collaborative language documentation," in *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 2014, pp. 1–5.

[16] V. Mitra, D. Vergyri, and H. Franco, "Unsupervised learning of acoustic units using autoencoders and kohonen nets," in *(to appear) Proceedings of Interspeech*, 2016.

[17] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, Kyoto, Japan, 2012.

[18] Zellig Harris, "From phoneme to morpheme," *Language*, vol. 31, pp. 190–222, 1995.

[19] Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater, "Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models," in *Advances in Neural Information Processing Systems*, 2006, pp. 641–648.

[20] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, pp. 21–54, 2009.

[21] Benjamin Borshinger and Mark Johnson, "Exploring the role of stress in Bayesian word segmentation using adaptor grammars," *Transaction of ACL*, vol. 2, pp. 93–104, 2014.