A SUMMARY OF THE 2012 JHU CLSP WORKSHOP ON ZERO RESOURCE SPEECH TECHNOLOGIES AND MODELS OF EARLY LANGUAGE ACQUISITION

Aren Jansen,[1] *Emmanuel Dupoux,*[2] *Sharon Goldwater,* [3] *Mark Johnson,*[4] *Sanjeev Khudanpur,*[1] *Kenneth Church,*[5] *Naomi Feldman,*[6] *Hynek Hermansky,*[1] *Florian Metze,*[7] *Richard Rose,*[8] *Mike Seltzer,*[9] *Pascal Clark*[1]*, Ian McGraw,*[10] *Balakrishnan Varadarajan,*[11] *Erin Bennett,*[6] *Benjamin Borschinger,*[4] *Justin Chiu,*[7] *Ewan Dunbar,*[6] *Abdellah Fourtassi,*[12] *David Harwath,*[10] *Chia-ying Lee,*[10] *Keith Levin,*[1] *Atta Norouzian,*[8] *Vijayaditya Peddinti,*[1] *Rachael Richardson,*[6] *Thomas Schatz,*[12] *Samuel Thomas*[1]

[1]Johns Hopkins CLSP/HLTCOE, [2]EHESS France, [3]U. Edinburgh, [4]Macquarie University, [5]IBM Research, [6]U. Maryland, [7]Carnegie Mellon University, [8]McGill University, [9]Microsoft Research, [10]MIT CSAIL, [11]Google, [12]ENS/CNRS France

ABSTRACT

We summarize the accomplishments of a multi-disciplinary workshop exploring the computational and scientific issues surrounding zero resource (unsupervised) speech technologies and related models of early language acquisition. Centered around the tasks of phonetic and lexical discovery, we consider unified evaluation metrics, present two new approaches for improving speaker independence in the absence of supervision, and evaluate the application of Bayesian word segmentation algorithms to automatic subword unit tokenizations. Finally, we present two strategies for integrating zero resource techniques into supervised settings, demonstrating the potential of unsupervised methods to improve mainstream technologies.

Index Terms— zero resource, speech recognition, language acquisition, Bayesian word segmentation, speaker independence

1. INTRODUCTION

Zero resource speech technologies operate without the expert-provided linguistic knowledge that standard recognition systems rely on—transcribed speech, language models, and pronunciation dictionaries. A robust zero-resource system must instead discover this linguistic knowledge from speech audio automatically. The problem is similar to that facing human infants, who must specialize their speech perception and production systems to their native language (though perhaps with help from other sensory modalities). This has led to the emergence of parallel scientific and engineering communities working towards strikingly similar research objectives—to understand how linguistic structure can be discovered from speech. The workshop brought together researchers in speech recognition, computational linguistics, and cognitive science in order to encourage dialog and extend and integrate existing techniques for linguistic discovery. Moreover, we sought to identify common evaluation frameworks to benchmark efforts across disciplines and provide useful predictors for technological success that also have an obvious path for application in human studies.

The two core technological and scientific building blocks of interest in the workshop were phonetic and lexical discovery. Phonetic discovery, also known as fully unsupervised acoustic model training in the speech community, is the process of automatically identifying the categorical subword inventory and relating it to the underlying acoustics. Existing methods rely on an acoustic feature representation, a notion of distance in that feature space, and an unsupervised learning algorithm. Speaker independence remains a major stumbling block [1] and improving it can be tackled in any of these three components. Given limited success of core recognition architectures in the zero resource setting, several alternative acoustic front-ends and unsupervised acoustic models have been proposed in recent years [2, 3, 4, 5, 1, 6, 7, 8, 9, 10], though there has been limited effort to evaluate these methods in a systematic way. Lexical discovery is the process of automatically identifying meaningful word-sized units from speech. Speech recognition researchers have developed discovery systems that search for repeated acoustic patterns, treating the remainder of the corpus as background [2, 3]. Most cognitive models of lexical discovery, on the other hand, perform a complete *word segmentation*, attempting to identify the boundaries between every word [11, 12, 13, 14, 15]. While complete segmentation is more desirable, existing segmentation algorithms require as input a subword unit tokenization that in the zero resource setting requires a phonetic discovery module. However, it remains an open question whether a categorical subword structure should be prerequisite.

To date these phonetic and lexical discovery tasks have been mostly considered in isolation. However, it has been recently demonstrated that even imperfect subword units can improve word discovery and detection [1], while an incomplete lexicon can aid subword unit discovery [16, 1, 17]. Moreover, cognitive science studies suggest that infants learn by simultaneously refining their subword units and their lexicon [18], indicating an integrated solution is likely necessary. Ultimately we would like to use insights from both of the speech and cognitive traditions to develop methods that can work from acoustics but also incorporate some of the top-down information that has been shown to help in cognitive models [19, 20, 21]. Still, the limited cross-disciplinary collaboration has failed to produce an evaluation of even the most basic integration strategies. With these issues in mind, the workshop efforts were divided into three interrelated subtopics: (i) devising unified evaluation metrics, (ii) improving and evaluating speaker independence of features and unsupervised acoustic models, and (iii) evaluating Bayesian word segmentation algorithms on noisy tokenizations automatically extracted from real speech data. In parallel, a fourth subteam explored strategies for integrating zero resource techniques into mainstream speech technologies. Results and background are summarized in the following sections. The interested reader can find workshop presentation videos on the web [22].

2. CROSS-DISCIPLINARY EVALUATION CRITERIA

We restricted our pursuit of a common evaluation framework to the task of phonetic discovery. We took as a starting point a recently proposed zero resource evaluation metric [23] that attempts to de-

termine how well various speech representations enable discrimination between word example pairs having the same or different type. This evaluation, referred to below as the same-different task, may be applied to both vector time series representations of speech (e.g. acoustic features or acoustic model posteriorgrams) and tokenizations (1-best subword unit decodes). This provides a unified means to evaluate representational quality regardless of data type, which by itself already goes a long way in unifying the evaluation of multiple computational strategies.

For a given multi-speaker corpus, the evaluation requires a presegmented collection of word examples (11k in our experiments below) provided by a forced alignment of the transcript. For all pairs of word examples (some 60 million), we compute a word-level dissimilarity, which we define as normalized dynamic time warping (DTW) distance for vector representations and as string edit distance for subword unit tokenizations. This dissimilarity is used to discriminate between same- and different-type pairs, where we characterize performance with average precision (AP). High AP indicates that the given representation is consistent across speakers. For supervised acoustic models, [23] demonstrated perfect correlation between AP and phone recognition error rate; given this AP can also be computed for features and unsupervised models, it is an ideal proxy.

The remaining question is whether this is also a natural evaluation metric for testing linguistic or scientific plausibility of a given phonetic discovery procedure. One appealing property is that such word level discrimination tasks can be conducted with human subjects, including infants. In addition, the scores on the same-different task can be analyzed for individual word type pairs, enabling more refined diagnostics of the front-ends. Further analysis was performed on the TIMIT same-different evaluation word set defined in [10]. We found it contains over 100k near-minimal pairs (edit distance less than 50% of string length), enabling secondary evaluation of representational performance on a wide array of phonetic contrasts. In particular, we successfully reconstructed phoneme- and feature-based confusion matrices using linear regression models of type-restricted same-different scores.

3. SPEAKER INDEPENDENCE OF ACOUSTIC FEATURES AND UNSUPERVISED MODELS

Recent zero resource efforts [2, 3, 4, 5] clearly demonstrate that front-ends that work best with supervised back-ends are not optimal for unsupervised learning. Likewise, stripped of any guidance from word transcripts and a pronunciation dictionary, the normal expectation-maximization training procedures for Gaussian mixture-based acoustic models are no longer capable of identifying speaker independent phonetic categories in a purely bottom-up fashion [24]. With these considerations in mind, an explicit goal of the workshop was to evaluate a variety of acoustic front-ends and unsupervised acoustic modeling strategies, both in isolation and in combination, for suitability in downstream zero resource technologies. In doing so, we have taken the initial steps in benchmarking competing approaches being developed across several institutions. Moreover, workshop participants identified two new approaches to improving representational speaker independence, which are summarized below and followed by the evaluation results.

3.1. Spectral Smoothing and Top-Down Lexical Constraints

A main source of speaker variation is the formant pattern changes that result from vocal cavity variation. Hermansky and Broad [25] demonstrated that the front cavity of the vocal tract, which is relatively invariant under changing vocal tract lengths, determines the phonetic value of the vowel and can be decoded by identifying the F2′ (effective second formant). It was shown that the F2′ value is closely tracked by the peaks of lower order perceptual linear prediction (PLP) spectral estimates. However even after PLP smoothing of the spectra, the cepstral transformation and subsequent mean variance normalization can still increase the effect of higher order cepstra on distance measures. Thus, team members considered two signal processing derived methods for improving front-end speaker independence: (i) reducing PLP model order O (default is 12) and (ii) cepstral truncation to D components (default is 13).

We also investigated the use of top-down lexical constraints to complement bottom-up unsupervised acoustic model training algorithms, e.g. [4, 7, 8, 9, 5]. In traditional supervised settings, lexical constraints come in the form of orthographic transcripts and pronunciation dictionaries. In the zero resource setting, we can fall back onto spoken term discovery algorithms that perform exhaustive searches through large corpora for word repetitions using nothing but the raw acoustic features as input [2]. While these algorithms do not provide the identity of the discovered items, they provide evidence that repetitions should have similar underlying subword unit sequences. This is a much weaker form of supervision, but it comes at little cost. The approach considered in the workshop, described in detail in [24], consists of four steps: (1) training a 1024-component Gaussian mixture model (GMM) on a large sample of in-domain audio, which serves as a sort of universal background model (UBM) for all speech sounds; (2) running a spoken term discovery system across the speech collection to produce a collection of word or phrase segment pairs and compute UBM posteriorgrams for each segment; (3) performing a DTW alignment of the acoustic frames of each word segment pair and use the frame-level correspondences to construct a similarity matrix over UBM components; and (4) partitioning the UBM Gaussian components with spectral clustering [26] and using each subset to define a subword unit GMM.

3.2. Same-Different Evaluation Results

We computed same-different task performance for several acoustic front-ends, including mel frequency cepstral coefficients (MFCC), PLP, frequency domain linear prediction (FDLP) [27], and nonlinear intrinsic spectral analysis (ISA) [10]. All acoustic features included velocities/accelerations and were globally mean and variance normalized. On the unsupervised acoustic model side, we considered output posteriorgrams from the standard bottom-up GMM-based UBM [4] both with and without the weak top-down constraints described above, as well as posteriors generated by the nonparametric Bayesian (NP Bayes) training procedure described in [9]. Finally, we provide task performance for English phonetic posteriorgrams from neural network (NN) acoustic models, serving as a supervised performance ceiling. We considered the same-different evaluation of Sec. 2 for both Switchboard and TIMIT (see [23] and [10] for details). Workshop time limitations precluded evaluation of all front-end and model combinations, while scalability constraints limited ISA and NP Bayes experiments to TIMIT.

Table 1 lists the Switchboard same-different evaluation performance for several features and model posteriorgrams, along with the DTW frame-level metric. First, we found that halving both the PLP model order and number of cepstral coefficients improved same-different task performance by 20% relative. For FDLP, which includes gain normalization but no linear predictive spectral smoothing, simply truncating the cepstrum to five dimensions produced a

Table 1. Switchboard same-different evaluation results.

Representation	Metric	AP
FDLP	cosine	0.215
PLP	cosine	0.177
MFCC	cosine	0.191
Truncated FDLP ($D = 5$)	cosine	0.257
Truncated PLP ($O = 7$, $D = 6$)	cosine	0.212
PLP + GMM-UBM + Top-down, 100 units	symm KL	0.286
PLP + GMM-UBM + Top-down, 50 units	symm KL	0.238
PLP + GMM-UBM, 100 units	symm KL	0.196
PLP + GMM-UBM, 50 units	symm KL	0.151
English NN Posteriorgrams, 100 hr	symm KL	0.516
English NN Posteriorgrams, 10 hr	symm KL	0.439

Table 2. TIMIT same-different evaluation results.

Representation	Metric	AP
ISA	cosine	0.496
PLP	cosine	0.348
MFCC	cosine	0.338
MFCC + NP Bayes, 507 units	symm KL	0.445
MFCC + GMM-UBM, 507 units	symm KL	0.271
MFCC + GMM-UBM, 50 units	symm KL	0.236
ISA + NP Bayes, 474 units	symm KL	0.464
ISA + GMM-UBM, 507 units	symm KL	0.447
ISA + GMM-UBM, 50 units	symm KL	0.332
English NN Posteriorgrams	symm KL	0.846

similar relative gain. Next, we found that GMM-based unsupervised models are unable to substantially outperform the raw features themselves, indicating the need for some form of constraint to produce speaker independence. We observe that the weak top-down constraint mechanism described in 3.1 provides up to 57% relative improvement over bottom-up training alone, providing more evidence for the promise of multi-level integration.

Table 2 lists the TIMIT evaluation results. Previous work [10] demonstrated substantial improvements when using ISA instead of standard PLP and MFCC features. However, ISA has no explicit categorical structure, so we evaluated its combination with unsupervised acoustic models. In combination with GMM-based models, we found ISA to provide a large relative improvement over the MFCC + GMM-UBM counterpart. Moreover, with MFCC input, we found a drastic improvement when using the nonparametric Bayesian HMM-GMM techniques over simple expectation maximization estimation of GMMs. The ISA/NP Bayes combination produced the best acoustic model performance to date for this task. Still, the strong performance of ISA alone raises the issue of whether a categorical subword structure is even necessary to discover speaker independent lexical structure. Either way, while we have recovered 60% of the supervised performance ceiling, there is much more to be understood.

4. BAYESIAN WORD SEGMENTATION OF AUTOMATIC SUBWORD UNIT TOKENIZATIONS

Nearly all word segmentation models [11, 12, 13, 14, 15] have been evaluated on manual phonemic tokenizations of the speech rather than deriving them automatically from the acoustic stream. Acoustic model provided tokenizations introduce substantial noise into the system, which only increases as we move from supervised subword models to the unsupervised ones described above. As a first exploratory step on the road to a fully integrated lexical and phonetic discovery model, we evaluated the performance of three word segmentation models applied to subword tokenizations produced by both supervised and unsupervised acoustic models.

4.1. Models

We experimented with three recent nonparametric Bayesian models of word segmentation. All assume a generative process in which a sequence of words is generated, and then the boundaries between these words are removed to create the observed unsegmented sequence of phones. For inference, all models use Markov Chain Monte Carlo algorithms, which produce samples from the posterior distribution of segmentations given the observed corpus. The simplest of the three models is the Dirichlet process (DP) model [19], which assumes that the latent sequence of words was generated using a unigram model. The second model is the hierarchical Dirichlet process (HDP) model [19], which extends the DP model by assuming that the latent word sequence is generated from a bigram rather than a unigram model. Finally, we used the 'Colloc' model from [21], which is based on the Adaptor Grammar framework [28] and assumes that the latent word sequence is generated as a sequence of "collocations,", each consisting of one or more words. All three models compute the probability of the first occurrence of a word using a unigram phone model, and further occurrences of words/bigrams/collocations using (roughly) relative frequencies. Previous work using phonemic input [19, 21] showed that the Bigram and Colloc models produce much more accurate segmentations than the Unigram model, a result that was used to argue that the contextual information provided by word-level dependencies is important for successful segmentation. (In the same papers, the Colloc outperformed Bigram, largely due to more sophisticated inference methods.) One goal of the current study was to examine whether the word-level dependency claim also holds for noisier input.

4.2. Evaluation and Discussion

Our evaluation was performed on various tokenizations of the Switchboard corpus. As this differs from the Bernstein-Ratner/Brent corpus of child-directed speech [29, 12] that the models were previously tested on, we included as an upper baseline a *phonemic* transcription of Switchboard derived from the orthographic word transcripts and a pronunciation dictionary. We also ran the models on (i) phonetic transcriptions from the ICSI Switchboard Transcription Project [30], which have about 30-35% phone error rate in comparison to the phonemic transcription; (ii) the 1-best output of a supervised neural network (NN) phone recognizer (the 100-hr model used in Table 1) with a 50% phone error rate; (iii) a 1-best decode of unsupervised GMM-UBM acoustic model, one with 25 units and another with 50 units; and (iv) 1-best output of the top-down constrained model with 100 units (described in Sec. 3.1). The 1-best outputs were generated using a Viterbi decode using an ergodic HMM with one state per output unit, uniform transition probabilities, and emission likelihoods generated by the given subword models. For scoring, each subword unit token was assigned to reference word token using forced aligned word transcripts (boundary units went to word with majority overlap). Performance was measured using token F-score, which requires both boundaries of each word to be correct.

Table 3 lists the word segmentation token F-score for each tokenization and model combination. Results from the unigram and bigram models are from a single sample after 10k iterations using the hyperparameter values from [19]. Results from the collocation model (run for 1000 iterations) use hyperparameter inference and maximum marginal decoding, which combines information from the last 500 samples of a single MCMC run as in [21]. Although the phonemic Switchboard corpus is clearly more difficult than the corpus of child-directed speech, the models perform reasonably well on it, and (as in previous work) we find that the models with word-level

Table 3. Word segmentation token F-scores (%) for Bernstein-Ratner/Brent phonemes and various Switchboard tokenizations.

Tokenization	Unigram	Bigram	Colloc
Bernstein-Ratner/Brent Phonemes	54	71	86
Switchboard Phonemes	58	66	66
Switchboard Allophones	29	22	29
English NN, 100 hr	28	18	27
GMM-UBM, 25 units	4.4	2.8	3.7
GMM-UBM, 50 units	4.1	2.2	3.5
GMM-UBM + Top-down, 100 units	3.1	1.4	2.5

Table 4. Word recognition accuracies (%) with DNN features and semi-supervised training.

System	Accuracy
Conventional acoustic features (PLP) using 1 hr. of English training data (baseline)	28.8
DNN based features pre-trained using 31 hrs. German/Spanish and 1 hr English	41.0
Acoustic model self-training with DNN features	44.8

dependencies (Bigram and Colloc) perform better than the Unigram model. However, as the input becomes noisier, performance of all models degrades drastically, and the Unigram model performs better. This is likely due to a blowup in the number of distinct phone sequences for each word—in the phonemic representation, each word is always represented by the same phone sequence, yielding a total lexicon of 4107 unique units. But in the other tokenizations, the number of unique units ranges from 13,668 to 35,534 (in a perfect negative correlation with token F-score). Since the models use repeated phone sequences to identify words, they are left without much signal even for learning individual words, much less dependencies.

These results underscore the need for integrated models using both top-down and bottom-up information to simultaneously discover phones and words. The word segmentation part needs to allow for variability in the phonetic realization of words, but can also provide top-down pressure for sounds in similar contexts to be labeled as the same phone. Steps in this direction have been taken [16, 31, 32], but we are aware of no fully integrated system using speech data as input. It might also be possible to improve results on automatic tokenizations by identifying each token as, say, consonantal or vocalic. This permits more sophisticated sub-word models in the segmentation systems, such as a syllable model rather than a unigram phone model. For Switchboard phoneme, allophone and NN input (where the consonant/vowel distinction is available), we found that learning syllables as well as words and collocations ('Colloc-Syllable' model of [21]) improved performance by 8-14%.

5. AIDING DOWNSTREAM SUPERVISED APPLICATIONS

In addition to exploring algorithms for unsupervised phonetic and lexical discovery, we investigated the role zero-resource methods can play in downstream supervised tasks. Below we describe two such efforts in large vocabulary recognition and spoken term detection.

5.1. Data-driven Front-ends and Selective Self Supervision

Acoustic models for state-of-the-art automatic speech recognition (ASR) systems are typically trained using hundreds or thousands of hours of transcribed speech audio. In low resource scenarios, we seek multi-lingual and semi-supervised methods to leverage more easily acquired high-resource or untranscribed speech to improve our ASR performance with minimal cost. Two avenues were explored in the workshop: (i) a multi-lingual corpus was used to train a data-driven, language-invariant front-end for low-resource recognition; and (ii) untranscribed speech audio was automatically transcribed and used to augment to the labeled data for training, a procedure known as self-training [33, 34]. For (i), discriminative deep neural network (DNN) pre-training [35] was performed on a multilingual corpus consisting of 31 hours of German/Spanish and only a single hour of English. The output of an internal DNN layer was then used to generate so-called bottleneck features with which we train an English large vocabulary speech recognizer using only the same hour of transcribed English speech. For (ii), the recognizer from (i) was used to automatically transcribe an additional 14 hours of English speech. Utterances with high-confidence recognition outputs were selected to be fed back for recognizer re-training. Table 4 shows that both techniques can improve recognition accuracy up to 16% absolute (55% relative). For details, see [36].

5.2. Improving Keyword Search using Lexical Discovery

Spoken term detection (STD) systems provide an efficient means to search large speech corpora for user-specified query words and phrases. Lexical discovery systems can automatically identify words of possible interest [3], so we investigated their utility in improving the search results of a high resource STD system. The STD baseline was a hybrid two-pass index-based system [37], which produces a ranked list of speech intervals that likely contain the occurrences of the query term along with confidence scores. The word discovery system, described in [38], produces a list of speech interval pairs (also with a confidence score), that likely contain the same word or phrase, but whose identity is unknown. We used this unlabeled repetition information to help verify the STD system hypotheses in a graph-based approach. Results from the two systems define the graph nodes, the discovery confidence scores define edge weights, and the STD system hypotheses are rescored using random walks. Performance was evaluated on a course lecture search task using a set of 34 out-of-vocabulary query terms with 243 occurrences overall [37]. The figure-of-merit [37] of the original and fused systems were 38.7% and 43.7%, respectively—a 13% relative performance improvement from augmenting the core STD system with zero resource techniques. For details, see [39].

6. CONCLUSIONS

The research and development of zero resource speech technologies is in its infancy. By bringing together leading researchers in related areas, we were able to begin benchmarking state-of-the-art techniques and take the first steps towards multi-level component integration. The main conclusion is that there remains long way to go in bridging the gap between unsupervised and supervised performance, and simple combination of existing discovery techniques across levels of linguistic representation is not sufficient. The experiments of Sec. 4 clearly indicate there is equal room for improvement on both the acoustic processing and computational linguistics sides of the effort, though the independent pursuit of both is suboptimal. The speaker independence gains from including word-level top-down constraints on phonetic discovery (Sec. 3) add to the existing evidence that a unified multi-level discovery model will be required. Given the demonstrated success of non-parametric Bayesian models for phonetic and lexical discovery, the obvious next step is integration under this methodological umbrella. In the meantime, there is a clear avenue for impact, as even imperfect unsupervised techniques can be complementary to supervised systems.

7. REFERENCES

[1] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models," in *Interspeech*, 2011.

[2] A. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE T-ASLP*, vol. 16, no. 1, pp. 186–197, 2008.

[3] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Interspeech*, 2010.

[4] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *ASRU*, 2009.

[5] M.-H. Siu, H. Gish, S. Lowe, and A. Chan, "Unsupervised audio patterns discovery using HMM-based self-organized units," in *Proc. of Interspeech*, 2011.

[6] T.J. Hazen, M.-H. Siu, H. Gish, S. Lowe, and A. Chan, "Topic modeling for spoken documents using only phonetic information," in *Proc. of the ASRU*, 2011.

[7] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic subword units," in *ACL-08: HLT*, 2008.

[8] X. Anguera, "Speaker independent discriminant feature extraction for acoustic pattern matching," in *Proc. ICASSP*, 2012.

[9] C.-Y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. ACL*, 2012.

[10] A. Jansen, S. Thomas, and H. Hermansky, "Intrinsic spectral analysis for zero and high resource speech recognition," in *Proc. of Interspeech*, 2012.

[11] C. De Marcken, "The unsupervised acquisition of a lexicon from continuous speech," *arXiv preprint cmp-lg/9512002*, 1995.

[12] M. Brent, "An efficient, probabilistically sound algorithm for segmentation and word discovery," *Machine Learning*, vol. 34, pp. 71–105, 1999.

[13] M. Fleck, "Lexicalized phonotactic word segmentation," in *ACL*, Columbus, Ohio, 2008, pp. 130–138.

[14] M. Christiansen, J. Allen, and M. Seidenberg, "Learning to segment speech using multiple cues: A connectionist model," *Language and Cognitive Processes*, vol. 13, pp. 221–268, 1998.

[15] A. Venkataraman, "A statistical model for word discovery in transcribed speech," *Computational Linguistics*, vol. 27, no. 3, pp. 351–372, 2001.

[16] N.H. Feldman, T.L. Griffiths, and J.L. Morgan, "Learning phonetic categories by learning a lexicon," *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pp. 2208–2213, 2009.

[17] A. Martin, S. Peperkamp, and E. Dupoux, "Learning phonemes with a proto-lexicon," *Cognitive Science*, 2012.

[18] J.F. Werker and R.C. Tees, "Influences on infant speech processing: Toward a new synthesis," *Annual review of psychology*, vol. 50, no. 1, pp. 509–535, 1999.

[19] S. Goldwater, T. L. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[20] M. Johnson, "Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure," in *ACL*, Columbus, Ohio, 2008.

[21] M. Johnson and S. Goldwater, "Improving nonparametric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars," in *NAACL*, Boulder, Colorado, 2009.

[22] Johns Hopkins University, "http://www.clsp.jhu.edu/workshops/archive/ws-12/groups/mini-workshop," (online web resource).

[23] M. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. of ICASSP*, 2011.

[24] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *IEEE ICASSP*, 2013.

[25] H. Hermansky and D.J. Broad, "The effective second formant F2' and the vocal tract front-cavity," in *Proc. of ICASSP*, 1989, pp. 480–483.

[26] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[27] S. Thomas, S. Ganapathy, and H. Hermansky, "Phoneme recognition using spectral envelope and modulation frequency features," in *Proc. of ICASSP*, 2009.

[28] M. Johnson, T. L. Griffiths, and S. Goldwater, "Adaptor grammars: a framework for specifying compositional nonparametric Bayesian models," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., Cambridge, MA, 2007, MIT Press.

[29] N. Bernstein-Ratner, "The phonology of parent-child speech," in *Children's Language*, K. Nelson and A. van Kleeck, Eds., vol. 6. Erlbaum, Hillsdale, NJ, 1987.

[30] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *Proc. of ICSLP*, 1996.

[31] M. Elsner, S. Goldwater, and J. Eisenstein, "Bootstrapping a unified model of lexical and phonetic acquisition," in *ACL*, 2012.

[32] G. Neubig, M. Mimura, S. Mori, and T. Kawahara, "Bayesian learning of a language model from continuous speech," *IEICE Transactions on Information and Systems*, vol. E95-D, no. 2, pp. 614–625, 2 2012.

[33] G. Zavaliagkos, M. Siu, T. Colthurst, and J. Billa, "Using untranscribed training data to improve performance," in *ISCA ICSLP*, 1998.

[34] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *IEEE ICASSP*, 2006.

[35] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *IEEE ASRU*, 2011.

[36] S. Thomas, M.L. Seltzer, K.W. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *IEEE ICASSP*, 2013.

[37] A. Norouzian and R. Rose, "Facilitating open vocabulary spoken term detection using a multiple pass hybrid search algorithm," in *Proc. of ICASSP*, 2012.

[38] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU*, 2011.

[39] A. Norouzian, R. Rose, S. H. Ghalehjegh, and A Jansen, "Zero resource graph-based confidence measure estimation for open vocabulary spoken term detection," in *IEEE ICASSP*, 2013.