SUBMODULAR DATA SELECTION WITH ACOUSTIC AND PHONETIC FEATURES FOR AUTOMATIC SPEECH RECOGNITION

Chongjia Ni[1], *Lei Wang*[1], *Haibo Liu*[2], *Cheung-Chi Leung*[1], *Li Lu*[2], *and Bin Ma*[1]

[1] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[2] Tencent Inc., Beijing, P. R. China
{nicj, wangl, ccleung, mabin}@i2r.a-star.edu.sg {geneliu, adolphlu}@tencent.com

ABSTRACT

In this paper, we propose to use acoustic feature based submodular function optimization to select a subset of untranscribed data for manual transcription, and retrain the initial acoustic model with the additional transcribed data. The acoustic features are obtained from an unsupervised Gaussian mixture model. We also integrate the acoustic features with the phonetic features, which are obtained from an initial ASR system, in the submodular function. Submodular function optimization has been theoretically shown its near-optimal guarantee. We performed the experiments on 1000 hours of Mandarin mobile phone speech, in which 300 hours of initial data was for the training of an initial acoustic model. The experimental results show that the acoustic feature based approach, which does not rely on an initial ASR system, performs as well as the phonetic feature based approach. Moreover, there is complementary effect between the acoustic feature based and the phonetic feature based data selection. The submodular function with the combined features provides a relative 4.8% character error rate (CER) reduction over the corresponding ASR system using random selection. We also include the desired feature distribution obtained from a development set in a generalized function, but the improvement is insignificant.

Index Terms— Active learning, data selection, automatic speech recognition, submodular optimization

1. INTRODUCTION

A large amount of speech data can easily be obtained via telephone calls or voiced-based applications such as voice search and voice message. To effectively utilize the large amount of untranscribed data, semi-supervised approaches [1-6] have attracted researchers' attention. For example, confidence-based approach [3,4,6] was proposed to first decode the untranscribed utterances using an existing ASR system, and then select the utterances with high confidence scores. The selected utterances together with their decoding hypotheses are then used to update the initial acoustic model. However, the confidence-based approach might prefer to select the data which is close to the training data in terms of speaking styles, noise types and content. Consequently, it could restrict the diversity of the training data. To avoid such shortcoming, active learning techniques [4,6,7] were proposed to select a small portion of data for manual transcription. The advantages of doing so are: (1) We do not enforce the knowledge the initial system knows. Note that the initially learned knowledge can be prone to errors. (2) The diversity of the selected data can be improved [6-8].

This work emphasizes on the active learning techniques in the following ASR application scenario: given an initial ASR system and a large amount of untranscribed speech data, the objective is to identify a small portion of data for manual transcription and add the newly transcribed data to retrain the acoustic model. The selected data should provide maximum contribution to the performance of the ASR system.

To address the above data selection problem, different active learning techniques [9-15] have been examined. For example, confidence-based approach was applied to acoustic modeling [12-15], and the utterances with low confidence scores were selected for manual transcription. The low-confident utterances are considered to be not well modeled by the existing acoustic model, and they are usually distorted by noise, spoken with accents or inarticulate. Such data can augment the diversity of the training set. Also, based on an existing acoustic model, Yu *et al.* [6] proposed to select data by maximizing the lattice entropy reduction over the entire database. Alternatively, Wu *et al.* [10] considered the data distribution, and selected data uniformly according to the predefined target speech units such as phonemes and words. Similarly, Siohan [8] selected data according to the distribution of context-dependent HMM states in a development set. Itoh *et al.* [9] suggested that both informativeness and representativeness of the data should be assessed at the same time. However, there is no optimal guarantee in terms of the objective function being optimized.

To overcome the above limitation, submodular optimization was examined and applied to active data selection. In submodular data selection, much work investigated the data selection based on the diversity of either the phonetic or the acoustic information. Wei *et al.* utilized tri-phone as phonetic feature [16] in the submodular function to select a subset from the transcribed training data to build an acoustic model. Wei *et al.* used the string kernel submodular function based on hypothesized phonetic label to select a subset to build a phone recognizer [17]. Shinohara [18] used the tri-phone distribution in the submodular function closed to a desired (uniform) distribution. In acoustic only approaches, Lin *et al.* [19] proposed to use submodular active selection on a Fisher-kernel based graph over untranscribed utterances, hence the pairwise similarities between all the utterances were computed. In a later work, Wei *et al.* [20] used a two-layer of acoustic features in the feature-based submodular function for the data selection of untranscribed data for acoustic model training. Their experimental results on the TIMIT corpus showed that the subset selected from untranscribed data could perform as well as if the transcription was known.

Inspired by the above mentioned work [16-20], we propose to use the acoustic features obtained from an unsupervised Gaussian mixture model (GMM), and integrate the acoustic features with the phonetic features (obtained from an initial ASR system) in the submodular function for the previously mentioned ASR application scenario. Different from other feature based submodular functions, computation of the similarity between any two utterances is hence avoided. Moreover, we investigate the effect of the generalized submodular function (with a desired feature distribution) as in [18] on the performance of the newly trained acoustic model. In addition to using a uniform distribution, we use the desired feature distribution obtained from a development set.

2. BACKGROUND

Submodular functions have been examined and applied to speech data selection [16-20]. The concept of submodularity refers to one type of properties of set-valued function.

Suppose $f: 2^V \to \Re$ to be a set-valued function, where $V = \{u_1, u_2, \cdots, u_N\}$ represents a set of *N* speech utterances. The function f is submodular if for every $A \subseteq B \subseteq V$ and $s \in V \setminus B$,

$$f(B \cup \{s\}) - f(B) \le f(A \cup \{s\}) - f(A)\,. \qquad (1)$$

Submodularity means that the gain by adding an element into a smaller set should not be less than that by adding the element into a superset. A submodular function f is monotone non-decreasing if $f(A \cup \{s\}) - f(A) \ge 0$ for $\forall s \in V \setminus A, A \subseteq V$. A submodular function f is normalized if *f*(Ø)= 0.

For ASR application, a subset S of training data V that maximizes the objective function f at a constraint should be selected. That is,

$$\max_{S \subseteq V} \{ f(S) : c(S) \le K \} \qquad (2)$$

where $c(S) \le K$ is the constraint.

The optimal problem is NP hard, and while it is NP hard, it can be approximately solved by using a greedy forward-selection algorithm, which is near-optimal as guaranteed by theorems proved by Nemhauser *et al.* [21]. Moreover, the greedy algorithm likely provides the best solution obtained in polynomial time unless P=NP [22].

3. FEATURE BASED SUBMODULAR FUNCTION

The submodular function f may take various forms, and there are several different submodular functions proposed in the previous works [16-20]. In [19], $f_{fac}(S) = \sum_{i \in V} \max_{j \in S} w_{ij}$ is used as the submodular function, but it requires to compute the similarity w_{ij} between any two utterances. In [16,20], researchers examined the feature-based submodular function $f_{fea}(S) = \sum_{u \in U} g(m_u(S))$, where $m_u(S) = \sum_{s \in S} m_u(s)$ measures the degree of feature *u* in the subset S and $g(\bullet)$ is a monotone non-decreasing function. Hence it is a two-layer feature-based submodular function and it also avoids computing the similarity between pairwise utterances. However, it only considers either the phonetic or the acoustic features.

For an ASR application, the selected utterances should match those in the application domain. Let $P = \{p_u\}_{u \in U}$ be the probability distribution over the feature set U, which is often used to characterize the application domain, and can be estimated from a development set. The normalized function $\overline{m}_u(S) = \frac{m_u(S)}{\sum_{u \in U} m_u(S)}$ can be seen as a distribution over the feature set U, and $M = \{\overline{m_u}(S)\}_{u \in U}$ denotes the probability of distribution.

Consider the KL-divergence between the two distributions $D(P \| M)$, then the following equation can be obtained:

$$D(P \| M) = const. + \log\left(\sum_{u \in U} m_u(S)\right) - \sum_{u \in U} p_u \log(m_u(S))\,. \qquad (3)$$

Then, we define a set-valued function

$$\begin{aligned} f_{dev-matched-fea}(S) &= \log\left(\sum_{u \in U} m_u(S)\right) - D(P \| M) \\ &= \sum_{u \in U} p_u \log(m_u(S))\,. \qquad (4) \end{aligned}$$

The function $f_{dev-match-fea}(S)$ is a submodular function according to submodular optimization theory [23]. It can represent the combination of its quantity of S via its features, and the feature distribution is close to the distribution P. Note that when $P = \{p_u\}_{u \in U}$ is in a uniform distribution, Eq. (4) has its form similar to $f_{fac}(\bullet)$.

Term frequency-inverse document frequency (tf-idf), is one of the representation for a document in the vector space model, and is used to reflect how important a word is to a document in a corpus. It is often used in information retrieval and text mining. In ASR application, an utterance can be seen as a document, which can be represented by a tf-idf vector.

For a given utterance s and feature u, the value of $m_u(s)$ can be computed by using the tf-idf vector, That is, $m_u(s) = tf(u,s) \times idf(u)$.

3.1 Phonetic feature based function

To generate the features U of untranscribed utterances, an initial ASR system is used to decode the utterances into sequences of phonemes or phoneme states. When a relatively low CER can be obtained, it is reasonable that the best decoding phoneme or phoneme state path is selected as the phoneme representation for an utterance. Thus, when selecting the phoneme states or n-gram phoneme states as features, the value of $m_u(s)$ can be computed by using $m_{u_1}(s) = tf(u_1, s) \times idf(u_1)$, where $u_1 \in U_1$ is a phonetic feature, and U_1 is the phonetic feature set.

3.2 Acoustic feature based function

Gaussian mixture model (GMM) is widely used to capture the acoustic characteristics of utterances. In text-independent speaker recognition, GMM is used as a universal background model to capture the general speech characteristics of a population of speakers [24]. The model captures not only speaker variation but also environmental variation. Moreover, each Gaussian component in the model can represent a phoneme class sharing similar acoustic characteristics [25]. In zero-resource speech processing, GMM is used to derive robust unsupervised posterior features for audio-only spoken term detection [25], subword unit discovery [26,27] and topic segmentation [28]. In this paper, we use GMM to characterize the acoustic property of each utterance for selection.

The quality, signal-to-noise ratio (SNR), accent, and speaking rate of these speech data are not reflected in transcriptions. Therefore, it is insufficient for selecting utterances only based on their phonetic information. The acoustic property should be considered when selecting utterance. In this paper, Algorithm 1 is proposed to extract acoustic feature for data selection.

Algorithm 1: **Acoustic feature extraction for data selection**

Step 1: Extract the spectral features (MFCC or PLP) of all the utterances for selection, and train a GMM using the spectral features.

Step 2: Decode each utterance by using the GMM, and output the index of the Gaussian component with the maximum posterior probability along the frame sequence of the utterance.

Step 3: Compute the n-gram counts of each utterance based on the frame sequence of the Gaussian component indices.

Step 4: Compute the tf values of each utterance, and compute the idf value of each Gaussian component index n-gram by using all the utterances.

Step 5: Compute the tf-idf values of each utterance.

Based on Algorithm 1, the n-grams of Gaussian component indexes are selected as features, and the value of $m_u(s)$ can be computed by $m_{u_2}(s) = tf(u_2, s) \times idf(u_2)$, where $u_2 \in U_2$ is an acoustic feature, and U_2 is the acoustic feature set.

3.3 Submodular function fusing phonetic and acoustic features

In order to fuse the phonetic and acoustic features of the utterances, the following function is proposed to use and select utterances:

$$f_{fused-fea}(S) = \sum_{u_1 \in U_1} p_{u_1} g_1\left(m_{u_1}(S)\right) + \alpha \sum_{u_2 \in U_2} g_2\left(m_u(S)\right), \quad (5)$$

where $g_1(\bullet)$ and $g_2(\bullet)$ is a monotone non-decreasing function, U_1 is the phonetic feature set, U_2 is the acoustic feature set, $\alpha > 0$ is the weight used to trade-off between the phonetic and acoustic representations, $\{p_{u_i}, u_i \in U_1\}$ is phonetic feature distribution, and can be estimated from a development set, and $m_u(S) = \sum_{s \in S} m_u(s)$.

Eq. (5) is a submodular function according to submodular function optimization theory [23], and it considers both the phonetic and the acoustic features in speech data selection task.

4. EXPERIMENTS AND RESULTS

4.1 Experimental setup

To verify the proposed acoustic feature based submodular function and the fusion of phonetic and acoustic features, data selection experiments were conducted on 1000 hours of Mandarin mobile phone speech data. The data was collected at 2 phases: i) A subset of 300 hours was first collected and manually transcribed to train an initial acoustic model; ii) Another subset of 700 hours of speech data was collected for data selection purpose. The data in both subsets consists of read speech and spontaneous speech with ratios 55% and 45% respectively. The read speech utterances were collected from more than 3000 speakers in a quiet environment, and the prompts were phonetically-rich sentences. While the spontaneous speech utterances were recorded through a simulated messaging system in both quiet and noisy environments, there were no constraints on the content. All the 1000 hours of data was recorded using different iOS, Android and Windows OS mobile devices, and the data format was PCM with 16KHz sampling rate, 16 bits and mono channel.

The task was to select 100 hours of speech data from the 700 hour subset. The selected data was then combined with the 300 hour subset to train an acoustic model for ASR evaluation. An open test set of 10 hour spontaneous speech was collected to evaluate the ASR performance which was measured by character error rate (CER). In addition, a non-overlapping development set which consists of 60 hour spontaneous data was used to estimate the phoneme state distribution in $f_{dev-matched-fea}(\bullet)$ of Eq. (4). The following 6 data selection approaches were examined and compared:

- Random-selection: data was selected randomly.
- Confidence-based-selection: data was selected with the lowest confidence scores [12-15].
- Phonetic-feature-based-selection: data was selected using $f_{fea}(\bullet)$ as submodular function with the best decoding state as feature [16].
- Dev-matched-phonetic-feature-based-selection: data was selected using $f_{dev-matched-fea}(\bullet)$ as submodular function with the best decoding state as feature.
- Acoustic-feature-based-selection: data was selected using $f_{fea}(\bullet)$ as submodular function with n-gram Gaussian component index as feature.
- Fusing-phonetic-acoustic-based-selection: data was selected using $f_{fused-fea}(\bullet)$ as submodular function, and fusion of the best decoding state and n-gram Gaussian component index was used as feature.

In the experiments, 52-dimensional features, including 12-dimensional mel frequency cepstral coefficient (MFCC) and 1-dimensional pitch along with their 1st, 2nd, and 3rd derivatives, were used. The cross-word tri-phone models represented by 3-state left-to-right HMMs were trained using boosted MMI discriminative training criterion [29]. State-clustered tri-phone HMMs contained 6500 states, and each tied state was modeled by 32 Gaussian components. The tri-gram language model, which consisted of about 118 thousand words and 11 millions of n-gram entries, was used to evaluate different systems performance. In Algorithm 1, the number of mixture components was 4096, and 2-gram Gaussian component index was used. In Eq. (5), we set $\alpha = 0.2$ and $g_1(x) = g_2(x) = \log(\bullet)$.

4.2 Experimental results and analysis

There are two baseline systems in our experiments. The baseline system "Baseline-300h" is built by using the 300 hours of data collected at the first phase, and the baseline system "Baseline-300h + Random-selection" is built by using combining randomly selected 100 hours of data with the data collected at the first phase. Table 1 lists the results on the test set.

Table 1. Baseline systems testing results

System	CER(%)
Baseline-300h	22.3
Baseline-300h + Random-selection	20.8

From the Table 1, we can find that with the augment of training data, the system performance can improve.

Table 2. Different systems testing results

System	CER(%)
Baseline-300h + Confidence-based-selection	20.6
Baseline-300h + Phonetic-feature-based-selection	20.2
Baseline-300h + Dev-matched-phonetic-feature-based-selection	20.1
Baseline-300h + Acoustic-feature-based-selection	20.1
Baseline-300h + Fusing-phonetic-acoustic-based-selection	19.8

In Table 2, system "Baseline-300h + Confidence-based-selection" is built by using combining confidence based selected 100 hours of data with the first phrase collected 300 hours of data. In Table 2, other systems are built similar to "Baseline-300h + Confidence-based-selection" system.

When comparing Table 2 with Table 1, we can find that: (1) The confidence based data selection and the submodular based data selection are better than the random data selection. (2) The Dev-matched-phonetic-feature-based-selection is slightly better than Phonetic-feature-based-selection, but the difference is insignificant. Despite similar performance, we find that Dev-matched-phonetic-feature-based-selection select more utterances from the spontaneous speech subset (~90% of the selected 100 hour data) than Phonetic-feature-based-selection (~75%). (3) Only using Gaussian component index n-gram as utterance representation also shows positive effect, and it can help us to select useful utterance for ASR. In our companion paper, a similar acoustic feature based submodular approach also shows its effectiveness for unsupervised data selection in a keyword search task [30,31]. (4) Fusing the phonetic features and acoustic features can help us to select more useful utterances for ASR. When comparing with the 300 hour baseline system, there is an 11.2% relative CER reduction. When comparing the ASR built by using our proposed approach with the 400 hour baseline system, there is a 4.8% relative CER reduction. From the experimental results, we also find that there are complementary effects between phonetic feature based data selection and the acoustic feature based data selection. We also observe that the phonetic features and acoustic features select utterances quite differently in terms of the types of recording mobile phones and speakers.

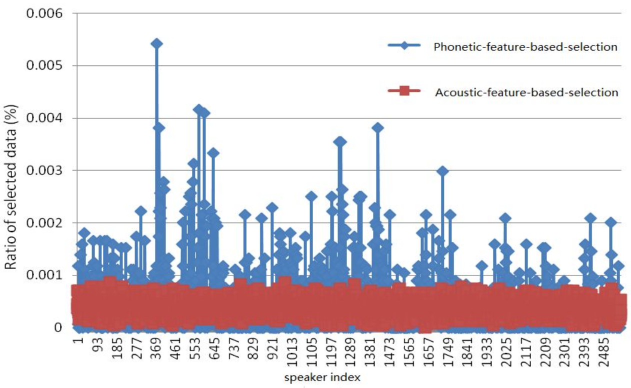

Fig. 1. Percentage of data selected from different speakers

Fig. 1 shows the percentage of data selected from different speakers in the spontaneous speech subset. In this subset, there is an equal amount of data from different speakers available for selection. From the figure, we find that the phonetic based approach chose the data from each speaker more evenly than the acoustic feature based approach. The phonetically-rich sentences in the data subset probably made the more even data selection among speakers. And the uneven inter-speaker variation probably made the more uneven data selection among speakers in the acoustic feature based approach. We believe that the complementary effect of the two data selection approaches leads to the selection of more useful utterances, which further improves the ASR performance.

5. CONCLUSION

In this paper, we propose to use an acoustic feature based submodular function for data selection. As other feature based submodular functions, the computation of the similarity between any two utterances is avoided. This feature based approach is feasible for the selection from a large amount of data. In our experiments, it is encouraging that the acoustic feature based approach, which does not requires an initial ASR system, performs as well as the phonetic feature based approach. Moreover, we propose to combine the acoustic and the phonetic features in the feature based submodular function. We find that there are complementary effects between acoustic feature based and phonetic feature based data selection. The submodular function with the combined features provides a relative 4.8% and 3.9% CER reduction over the corresponding ASR system using random selection and confidence based selection respectively. The GMM that we use to obtain the acoustic features probably captures both speaker and environment variation. In the future, we would study the effect of these two kinds of variation to the data selection separately, evaluate the data selection algorithms on deep neural network acoustic models, and study the effectiveness of our proposed approach when more training data is involved.

6. REFERENCES

[1] F. Wessel, and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," IEEE Trans. on Speech and Audio Processing. 2005,13(1):23-31.

[2] T. Kemp, A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," in Proc. Eurospeech 1999, pp. 2725-2728.

[3] D. Charlet, "Confidence-measure-driven Unsupervised Incremental Adaptation for HMM-based Speech Recognition," in Proc. ICASSP 2001, pp. 357-360.

[4] G. Tur, D. Hakkani-Tur and R. E. Shapire, "Combining Active and Semi-supervised Learning for Spoken Language Understanding," Speech Communication, 2005,45(2):171-186.

[5] X. Zhu, "Semi-supervised Learning Literature Survey," Computer Sciences Technical Report 1530, University of Wisconsin-Madison, 2005b.

[6] D. Yu, B. Varadarajan, L. Deng and A. Acero, "Active Learning and Semi-Supervised Learning for Speech Recognition: A Unified Framework Using the Global Entropy Reduction Maximization Criterion," Computer Speech and Language, 2010, 24(3): 433-444.

[7] B. Settles, "Active Learning Literature Survey," Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2010.

[8] O. Siohan, "Training Data Selection Based on Context-Dependent State Matching," in Proc. ICASSP 2014, pp. 3316-3319.

[9] N. Itoh, T. N. Sainath, D. N. Jiang, J. Zhou, and B. Ramabhadran, "N-Best Entropy Based Data Selection for Acoustic Modeling," in Proc. ICASSP 2012, pp. 4133-4136.

[10] Y. Wu, R. Zhang, and A. Rudnicky, "Data Selection for Speech Recognition," in Proc. ASRU 2007, pp. 562-565.

[11] Y. Hamanaka, K. Shinoda, T. Tsutaoka, S. Furui, T. Emori, and T. Koshinaka, "Committee-Based Active Learning for Speech Recognition," IEICE Trans. on Information and Systems, 2011, 10: 2015-2013.

[12] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active Learning for Automatic Speech Recognition," in Proc. ICASSP 2002, pp. 3904-3907.

[13] G. Riccardi and D. Hakkani-Tur, "Active and Unsupervised Learning for Automatic Speech Recognition," in Proc. Eurospeech 2003, pp. 1825-1828.

[14] G. Tur, R. E. Schapire, and D. Hakkani-Tur, "Active Learning for Spoken Language Understanding," in Proc. ICASSP 2003, pp. I-276-I-279.

[15] T. M. Kamm and G. G. L. Meyer, "Selective Sampling of Training Data for Speech Recognition," in Proc. Human Language Technology Conf., San Diego, CA, 2002.

[16] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels and J. Bilmes, "Submodular Subset Selection for Large-Scale Speech Training Data," in Proc. ICASSP 2014, pp. 3311- 3315.

[17] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Using Document Summarization Techniques for Speech Data Subset Selection," in Proc. NAACL/HLT-2013, pp. 721-726.

[18] Y. Shinohara, "A Submodular Optimization Approach to Sentence Set Selection," in Proc. ICASSP 2014, pp. 4140-4143.

[19] H. Lin and J. Bilmes, "How to Select a Good Training-data Subset for Transcription: Submodular Active Selection for Sequences," in Proc. Interspeech 2009, pp. 2859-2862.

[20] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Unsupervised Submodular Subset Selection for Speech Data," in Proc. ICASSP 2014, pp. 4107-4111.

[21] G. Nemhauser, L. Wolsey, and M. Fisher, "An Analysis of Approximations for Maximizing Submodular Set Function-I," Mathematical Programming,1978, 14(1):265-294.

[22] U. Feige, "A Threshold of ln n for Approximating Set Cover," Journal of the ACM (JACM), 1998,45(4):634-652.

[23] F. Bach, "Learning with Submodular Functions: A Convex Optimization Perspective," Foundations and Trends ® in Machine Learning," 2013,6(2-3): 145-373.

[24] T. Kinnunen, and H. Li, "An Overview of Text-independent Speaker Recognition: From Features to Supervectors," Speech Communication, 2010:52(1):12-40.

[25] Y. Zhang, "Unsupervised Speech Processing with Applications to Query-by-Example Spoken Term Detection," Ph.D Thesis.

[26] H. Wang, T. Lee, C.-C. Leung, B. Ma and H. Li, "Unsupervised Mining of Acoustic Subword Units With Segment-level Gaussian Posteriorgrams," in Proc. Interspeech 2013, pp.2297-2301.

[27] H. Wang, T. Lee, C.-C. Leung, B. Ma and H. Li, "Acoustic Segment Modeling with Spectral Clustering Methods," IEEE Trans. On Audio, Speech and Language Processing, 2015,23(2):264-277.

[28] L. Zhang, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Acoustic TextTiling for Story Segmentation of Spoken Documents," in Proc. ICASSP 2012, pp.5121-5124.

[29] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for Model and Feature Space Discriminative Training," in Proc. ICASSP 2008, pp. 4057-4060.

[30] C. Ni, C.-C. Leung, L. Wang, N. F. Chen and B. Ma, "Unsupervised Data Selection and Word Morph Mixed Language Model for Tamil Low Resource Spoken Keyword Spotting," in Proc. ICASSP 2015.

[31] N. F. Chen, C. Ni, I-Fan Chen, S. Sivadas, V. T. Pham, H. Xu, X. Xiao, T. S. Lau, S. Leow, B. P. Lim, C.-C. Leung, L. Wang, C.-H. Lee, A. Goh, E. S. Chng, B. Ma, H. Li, "Low-resource Keyword Search Strategies for Tamil," in Proc. ICASSP 2015.