INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

On the Linguistic Relevance of Speech Units Learned by Unsupervised Acoustic Modeling

Siyuan Feng, Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
{syfeng, tanlee}@ee.cuhk.edu.hk

Abstract

Unsupervised acoustic modeling is an important and challenging problem in spoken language technology development for low-resource languages. It aims at automatically learning a set of speech units from un-transcribed data. These learned units are expected to be related to fundamental linguistic units that constitute the concerned language. Formulated as a clustering problem, unsupervised acoustic modeling methods are often evaluated in terms of average purity or similar types of performance measures. They do not provide detailed insights on the fitness of individual learned units and the relation between them. This paper presents an investigation on the linguistic relevance of learned speech units based on Kullback-Leibler (KL) divergence. A symmetric KL divergence metric is used to measure the distance between each pair of learned unit and ground-truth phoneme of the target language. Experimental analysis on a multilingual database shows that KL divergence is consistent with purity in evaluating clustering results. The deviation between a learned unit and its closest ground-truth phoneme is comparable to the inherent variability of the phoneme. The learned speech units have a good coverage of linguistically defined phonemes. However, there are certain phonemes that can not be covered, for example, the retroflex final /er/ in Mandarin.

Index Terms: unsupervised acoustic modeling, clustering, KL divergence

1. Introduction

Acoustic modeling (AM) refers to the task of learning and representing the statistical relation between speech signals and the basic phonetic units of a specific language. It is the core of automatic speech recognition (ASR) technology. Conventionally acoustic model training is done in a supervised manner, i.e., the training speech must be accompanied by detailed transcription at word level. As state-of-the-art ASR systems are typically built with thousand hours of speech, preparing training transcription is considered an unwelcome task and numerous approaches have been attempted to achieve semi-supervised or lightly-supervised training [1, 2, 3, 4]. There are also application scenarios that transcribing speech is simply not possible, for example, when no writing system and no pronunciation lexicon are available for the target language. In recent years, there has been a growing research interest in unsupervised acoustic modeling, which assumes that only un-transcribed raw speech are available [5, 6, 7, 8, 9]. This is a challenging task with significant practical impact. Unsupervised acoustic modeling has been investigated mainly in applications related to low-resource languages [10], as well as in language identification [11] and topic modeling [12]. In the latest international conferences, zero-resource speech technology remains a hot topic of research [13].

Given a certain amount of un-transcribed speech, unsupervised acoustic modeling can be formulated as a segmentation-clustering problem [8]. The speech signal is divided into variable-length segments, which are subsequently clustered into a limited number of groups based on acoustic similarity. Each group of segments is assigned a label that expectedly represents a basic sound unit of the language. With the labeled segments, supervised training can be applied to establish the acoustic models, which could be based on GMM-HMM [14], DNN-HMM [15] or LSTM-RNN [16].

The above approach to unsupervised acoustic modeling has been investigated extensively. On unsupervised segmentation, Qiao *et al.* [17] presented a bottom-up hierarchical segmentation algorithm that exploits the spectral discontinuities at segment boundaries. Torbati *et al.* [18] introduced a phoneme segmentation method based on Bayesian HMM with hierarchical Dirichlet process (HDP) priors. Estevan *et al.* [19] applied the maximum-margin clustering algorithm in speech segmentation. Feng *et al.* [9] proposed to use multiple language-mismatched phone recognizers to make phonetically-informed hypotheses of segment boundaries. For segment clustering, Wang *et al.* [20] applied the spectral clustering approach with segment-level posterior features, and demonstrated superior performance as compared to vector quantization (VQ) [21], Gaussian labeling (GL) [22] and segmental GMM (SGMM) [23].

The efficacy of acoustic models is evaluated typically in terms of speech recognition performance, e.g., phoneme accuracy or word error rate. This is obviously not applicable to the case of unsupervised acoustic modeling, in which there are not predefined phonemes and words. Purity [20][9], normalized mutual information (NMI) [20] and average precision (AP) [6] are commonly used performance metrics for evaluating clustering algorithms. These metrics facilitate straightforward comparison of overall performance. They do not provide detailed insights on the fitness of individual clusters and the relation between the clusters. For the investigation of unsupervised acoustic modeling methods, one of the major concerns is about the linguistic relevance of the automatically learned speech units. For example, let us consider the clustering results produced by two different clustering algorithms (or the same algorithm with different parameter settings). In the first case, the degree of overlap between an automatically learned cluster and its closest ground-truth phoneme varies greatly from one cluster to another, whereas in the second case, the degrees of overlap are equal across all clusters. Although the two sets of clustering results may give the same purity value, their linguistic implications could be very different. In this study, we investigate the use of Kullback-Leibler (KL) divergence in the analysis of automatically learned speech units in unsupervised acoustic modeling. KL divergence is a statistics-based distance measure that can be used to model implicit speech variation related to phonetic context, pronunciation variation, speaker characteris-

http://dx.doi.org/10.21437/Interspeech.2017-300

tics, etc. It was successfully applied to ASR acoustic modeling [24][25], training data selection [26], cross-lingual TTS [27] and voice conversion [28]. Motivated by the successful application of KL divergence metric in measuring phonetic distortion and performing senone mapping [27][28], we are interested in its effectiveness in analyzing the linguistic relevance of speech units automatically learned from an unknown language.

2. Unsupervised Acoustic Modeling Framework

The basic framework of unsupervised acoustic modeling consists of three stages: unsupervised segmentation, segment clustering and statistical modeling [9]. Speech utterances of the target language are first divided into segments of variable length based on the hypothesized segment boundaries that are generated by a few language-mismatched phone recognizers. A spectral clustering algorithm is applied to group the speech segments into a prescribed number of clusters. Each cluster of segments is given a label. The segment labels are regarded as phone-level transcriptions, with which supervised training of acoustic models can be carried out. The target language mentioned above is generally a low-resource language. However, in the present study, the target languages being modeled are commonly regarded as resource-rich languages. This is to facilitate the analysis of clustering results, which requires the availability of ground-truth transcription and time alignment.

The features extracted for segment clustering are segment-level posteriors. They are derived from the frame-level posteriors produced by the language-mismatched phone recognizers. Let $\boldsymbol{x_t}$ be the acoustic observation of the t-th frame. The frame-level posterior feature vector is defined as,

$$\boldsymbol{q_t} = \begin{bmatrix} p(c_1|\boldsymbol{x_t}) \\ \vdots \\ p(c_m|\boldsymbol{x_t}) \\ \vdots \\ p(c_M|\boldsymbol{x_t}) \end{bmatrix}, t = 1, 2, \ldots, T, \tag{1}$$

where $\{c_1, c_2, \ldots, c_M\}$ denote M phoneme classes covered by multiple recognizers, $p(c_m|\boldsymbol{x_t})$ is the posterior probability of c_m given $\boldsymbol{x_t}$. By taking the average of frame-level posterior vectors in a hypothesized segment, the segment-level posterior vector is obtained as,

$$\boldsymbol{\hat{x}_k} = \frac{1}{e_k - b_k + 1} \sum_{t=b_k}^{e_k} \boldsymbol{\hat{q}_t}, k = 1, 2, \ldots, K, \tag{2}$$

where K is the number of segments, b_k and e_k denote the beginning and the end of the k-th segment. By segment clustering, the hypothesized segments are grouped into a prescribed number of clusters. Each cluster corresponds to an automatically learned speech unit, which is represented by the probability distribution of frame-level posterior features. Similarly, if the ground-truth time alignments of all phonemes are given, each phoneme can be represented by the probability distribution of frame-level posterior features.

Each language-mismatched phone recognizer serves as a nonlinear mapping between the acoustic space and the phonetic space. Through this mapping, irrelevant variations of acoustic features, e.g., speaker, gender, channel and environmental noise, are suppressed and/or normalized. It is assumed, with good reasons, that the phonetic spaces of different languages are substantially overlapped, since they are all based on the same mechanism of speech production. Compared with acoustic features, posterior features are more reliable in reflecting inherent phonetic characteristics of the target language.

3. Distance Between Speech Classes and Ground-Truth Phonemes

We propose a KL divergence-based method to measure the distance between each pair of automatically learned speech unit and ground-truth phoneme in the target language. The KL divergence, also called information divergence, or relative entropy, is a measure of the difference between two probability distributions [29]. For discrete probability distributions P and Q, the KL divergence is defined as [30],

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{3}$$

Equation (3) gives a non-symmetric measure. In this paper the following symmetric discrete form of KL divergence is adopted,

$$D_{KL}(P, Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{4}$$

$$= \sum_i (P(i) - Q(i)) \cdot \log \frac{P(i)}{Q(i)}. \tag{5}$$

The symmetric KL divergence is used to measure the distance between posterior feature distributions between each pair of automatically learned speech unit and ground-truth phoneme.

Let $\{g_1, g_2, \ldots, g_K\}$ denote K ground-truth phonemes. $\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_{L_k}}$ are the posterior probability vectors of L_k frames that are labeled as phoneme g_k according to ground-truth transcription. $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_{L_k}}\}$ is a subset of $\{\boldsymbol{q_1}, \boldsymbol{q_2}, \ldots, \boldsymbol{q_T}\}$. The centroid of g_k is computed as,

$$\overline{\boldsymbol{v^k}} = \frac{\sum_{i=1}^{L_k} \boldsymbol{v_i}}{L_k}. \tag{6}$$

$\overline{\boldsymbol{v^k}}$ is treated as the representative of g_k in the phonetic space. Let $\{u_1, u_2, \ldots, u_R\}$ denote R automatically learned speech units, $\{\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \ldots, \boldsymbol{\mu_{N_r}}\} \subset \{\boldsymbol{q_1}, \boldsymbol{q_2}, \ldots, \boldsymbol{q_T}\}$ are the posterior probability vectors of frames assigned to u_r ($r = 1, 2, \ldots, R$). The distance between u_r and g_k is defined as,

$$D(u_r, g_k) = \frac{\sum_{j=1}^{N_r} D_{KL}(\boldsymbol{\mu_j}, \overline{\boldsymbol{v^k}})}{N_r} \tag{7}$$

$$= \frac{\sum_{j=1}^{N_r} \sum_{m=1}^{M} (\mu_j(m) - \overline{v^k}(m)) \cdot \log \frac{\mu_j(m)}{\overline{v^k}(m)}}{N_r}. \tag{8}$$

Here, the distortion between two acoustic models u_r and g_k is measured by the KL divergence-based distance between posterior features in the phonetic space.

For each learned speech unit u_r, let $g_{k^*}(u_r)$ denote the closest ground-truth phoneme, where

$$k^* = \arg\min_k D(u_r, g_k). \tag{9}$$

Let $g_{k^*}(u_r)$ be abbreviated as $g^*(u_r)$. The distance between u_r and $g^*(u_r)$, denoted as $D^*(u_r)$, is computed by

$$D^*(u_r) = D(u_r, g^*(u_r)). \qquad (10)$$

Subsequently, the distance between u_r and the second closest phoneme $g_{k^{**}}(u_r)$ (abbreviated as $g^{**}(u_r)$) is calculated by

$$k^{**} = \arg\min_{k \neq k^*} D(u_r, g_k). \qquad (11)$$

The distance $D(u_r, g^{**}(u_r))$ is denoted as $D^{**}(u_r)$. The discrimination capability of the cluster u_r can be measured by

$$\Delta D^*(u_r) = |D^{**}(u_r) - D^*(u_r)|. \qquad (12)$$

A small value of $D^*(u_r)$ means that the automatically learned unit matches well with one of the ground-truth phonemes. Meanwhile, a large value of $\Delta D^*(u_r)$ indicates that $g^*(u_r)$ is discriminatively mapped to u_r.

The distance measure can be further extended to evaluating inherent variability of each ground-truth phoneme. For the phoneme g_k, the inherent variability is obtained as $\widetilde{D}(g_k)$,

$$\widetilde{D}(g_k) = \frac{\sum\limits_{j=1}^{L_k} D_{KL}(\boldsymbol{v_j}, \overline{\boldsymbol{v^k}})}{L_k}. \qquad (13)$$

A small value of $\widetilde{D}(g_k)$ indicates that the acoustic-phonetic properties of g_k are highly consistent in the training speech. It must be noted that $D^*(u_r)$ computed for learned speech units and $\widetilde{D}(g_{k^*})$ for ground-truth phonemes are comparable, as both of them measure the deviation from a class of posterior feature vectors to the centroid of the same phoneme class g_{k^*}, in the same phonetic space. $\widetilde{D}(g_{k^*})$ is calculated from ground-truth transcription, and independent of the clustering results. Therefore $\widetilde{D}(g_{k^*})$ could be a good reference for $D^*(u_r)$.

The KL divergence metric in this paper is not only applicable to posterior features extracted from phone recognizers, but also to conventional spectral features like MFCCs, or neural network bottle-neck features (BNFs).

4. Experimental Design

An experimental system of unsupervised acoustic modeling is established to generate automatically learned speech units from un-transcribed speech data of a target language. The symmetric KL divergence is used to analyze the linguistic relevance of these learned units with respect to the ground-truth phonemes.

4.1. Databases

Spontaneous story-telling speech from the OGI Multi-language Telephone Speech Corpus (OGI-MTS) [31] are used in our experiments. There are five target languages involved: German (GE), Hindi (HI), Japanese (JA), Mandarin (MA) and Spanish (SP). The corpus provides manual time alignment at phoneme level. The amount of speech data (in hours) and the number of labeled phonemes (including silence) for each language are summarized as in Table 1.

Table 1: *Multi-lingual speech data from the OGI-MTS corpus*

Language:	GE	HI	JA	MA	SP
Data size:	1.31	0.95	0.86	0.57	1.46
# Phonemes:	43	46	29	44	38

4.2. Automatically Learned Speech Units

Implementation of the unsupervised acoustic modeling framework in this work is based on approaches in [9]. Four language-mismatched phone recognizers are used as feature extractors to generate frame-level posterior features. They are Czech (CZ), Hungarian (HU), Russian (RU) and Cantonese (CT) phone recognizers. The CZ, HU and RU recognizers were developed and made publicly available by Brno University of Technology [32]. The number of phonemes being modeled are 45, 61 and 52, respectively. The CT recognizer is trained with the CUSENT database, which was developed by The Chinese University of Hong Kong to support general ASR applications [33]. The number of phonemes in the CT recognizer is 73.

A spectral clustering algorithm is applied to cluster the segments into R clusters, where $R = 50, 60, 70, 80, 90$ in our experiments. Each of these clusters is assigned a label, which denotes one of the learned speech units. By assigning cluster labels to all segments in an input utterance, we essentially obtain a time-aligned transcription for the utterance.

The similarity between automatically learned transcription and ground-truth transcription can be quantified by the purity measure. Let R' be the number of ground-truth phonemes, $n_{r,r'}$ denote the number of frames assigned to the r-th cluster and labeled as the r'-th ground-truth phoneme. The overall purity is defined as,

$$\text{purity} = \frac{\sum\limits_{r=1}^{R} \max\limits_{r' \in \{1,2,\ldots,R'\}} n_{r,r'}}{\sum\limits_{r=1}^{R} \sum\limits_{r'=1}^{R'} n_{r,r'}}. \qquad (14)$$

The purity values for each target language with R ranging from 50 to 90 are shown as in Table 2.

Table 2: *Purity for the five target languages*

	$R = 50$	$R = 60$	$R = 70$	$R = 80$	$R = 90$	Average
GE	0.428	0.433	0.439	0.437	0.437	0.43
HI	0.494	0.499	0.492	0.489	0.485	0.49
JA	0.545	0.556	0.554	0.543	0.540	0.55
MA	0.414	0.425	0.434	0.426	0.426	0.42
SP	0.556	0.586	0.576	0.568	0.573	0.57

5. Results and Analysis

The relation between automatically learned speech units and ground-truth phonemes is analyzed for each target language, based on the clustering results and the ground-truth time alignment provided in the corpus. For each learned speech unit u_r, Equations (7), (9) and (10) are used to determine its closest ground-truth phoneme $g^*(u_r)$ and the distance $D^*(u_r)$ between them. Equation (11) is used to determine the second closest ground-truth phoneme $g^{**}(u_r)$ and the distance $D^{**}(u_r)$. For each ground-truth phoneme g_k, Equation (13) is used to calculate the inherent variability $\widetilde{D}(g_k)$. The average values of $D^*(u_r)$, $D^{**}(u_r)$ and $\widetilde{D}(g_k)$ are summarized in Table 3. As seen from Table 2 & 3, both the purity values and the average KL divergence are not sensitive to the cluster R.

Figure 1 compares the average values of $\Delta D^*(u_r)$ and the purity values for the five target languages. From Table 3 and Figure 1, the following observations are made:

(a) $\overline{D^*(u_r)}$ is smaller than or approximately equal to $\overline{\widetilde{D}(g_k)}$ for all target languages. In other words, the deviation

Table 3: $\overline{D^*(u_r)}/\overline{D^{**}(u_r)}$ *and* $\overline{\widetilde{D}(g_k)}$

	$R=50$	$R=60$	$R=70$	$R=80$	$R=90$	$\overline{\widetilde{D}(g_k)}$
GE	27.4/30.1	27.3/30.0	27.3/30.1	27.5/30.3	27.4/30.3	27.8
HI	27.6/31.5	27.6/31.6	27.8/31.7	28.0/31.7	28.1/31.8	28.5
JA	28.5/34.1	28.0/33.2	28.1/33.2	28.3/33.4	28.6/34.1	27.9
MA	29.3/31.2	29.3/31.5	29.0/31.0	29.5/31.6	29.4/31.3	29.1
SP	28.5/31.9	27.8/32.2	27.9/32.3	28.1/32.5	28.0/32.4	28.0

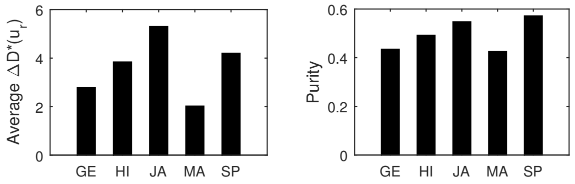

Figure 1: *Average values of* $\Delta D^*(u_r)$ *and purity values*

between an automatically learned speech unit and its closest ground-truth phoneme is comparable to, if not smaller than, the inherent variability of the phoneme itself. In fact, the ground-truth phonemes are labeled based on auditory perception of linguistic experts, whereas the learning of speech units, as well as the proposed KL divergence metric, is totally data-driven.

(b) The average values of $\Delta D^*(u_r)$ for different languages have the same trend as the purity values. This observation is consistent with our expectation, as a larger $\Delta D^*(u_r)$ implies that the automatically learned speech unit is mapped to its closest ground-truth phoneme with a higher confidence, thus naturally leads to a higher purity.

We are interested to understand more about the phonetic coverage of automatically learned speech units. Each learned unit corresponds to one best-matching ground-truth phoneme based on Equation (9). If a ground-truth phoneme fails to be selected as the best-matching phoneme for any of the learned units, it is regarded as not being covered. Table 4 shows the counts of uncovered vowels and consonants for each target language for $R = 90$. It can be seen that most of the linguistically-defined phonemes could be covered in the process of unsupervised acoustic modeling. Particularly in the case of Japanese, all phonemes are covered. However, there are quite a few phonemes of Mandarin that are not covered by the learned units. The missing vowels and consonants are {/aa/ (/a/)[1], /er/ (/ɚ/)} and {/kh/ (/kʰ/), /ph/ (/pʰ/), /r/ (/ɻ/), /tH/ (/tʰ/)}, respectively. It is interesting to see that the majority of missing consonants are unvoiced plosives. These consonants have strong transitory characteristics, i.e., rapidly changing spectral properties. In the segmentation process of unsupervised acoustic modeling, it is assumed that individual frames in the same segment have similar spectral properties. This assumption is not valid for transitory phonemes. Similarly, the missing vowel /er/, known as Erhuayin in Mandarin, also has transitory properties. This inspires us to investigate alternative features and segment representation which could capture trajectory characteristics of phonemes.

Table 4: *Uncovered/total No. vowels and consonants (*$R = 90$*)*

	GE	HI	JA	MA	SP
Vowels	1/18	0/13	0/7	2/17	1/11
Consonants	1/24	3/32	0/21	4/26	1/27

It is not expected that the vowel /aa/ is missed. Although /aa/ is not selected as the closest phoneme to any of the learned units, it is actually identified as the second closest phoneme to two different learned units. These two units are corresponded to /ae/ (/a/) and /aw/ (/au/) according to the KL divergence. In the transcription of Mandarin speech in OGI-MTS database, /aa/ is used to label the vowel nucleus in the Pinyin Finals /a/ and /ang/, while /ae/ is used to label the vowel nucleus in the Pinyin Final /an/ [34]. The two vowel nuclei are actually very similar in articulation. From this perspective, /aa/ is actually not a missing phoneme.

It must be noted that the identities of uncovered ground-truth phonemes depend also on experimental configurations, such as initialization of clustering, cluster number, etc.

For some of the learned units, the value of $D^{**}(u_r)$ is nearly the same as $D^*(u_r)$. In other words, such a learned unit matches equally well with two different phonemes. This kind of confusion can be alleviated with a large R. Table 5 gives an example of confusion between a few Japanese vowels. With $R = 50$, /uw/ (/ɯ/) and /iy/ (/i/) are the closest and second closest phonemes to cluster 33. Similar observation can be made on /uw/ and /ey/ (/e/) to cluster 41. When R is increased to 90, the confusion is significantly alleviated. A larger R leads to smaller-size as well as finer clusters, therefore the learned clusters containing segments of multiple ground-truth phonemes tend to split and form linguistically more explicit speech units.

Table 5: *Speech units mapped to /uw/ with* $R = 50$ *and* 90

# Clusters R		50		90			
Cluster label r		33	41	25	49	29	35
$D^*(u_r)$	/uw/	**33.9**	**35.9**	29.1	33.1	29.7	30.4
$D^{**}(u_r)$	/iy/	**34.1**	—	34.6	35.3	—	—
	/ey/	—	**35.9**	—	—	35.0	31.1

6. Conclusions

This paper presents a study on the linguistic relevance of speech units learned by unsupervised acoustic modeling. A symmetric KL divergence metric is defined and used to measure the distance between each pair of learned unit and ground-truth phoneme of the target language. Experimental results show that KL divergence is consistent with purity in evaluating clustering results. The deviation between a learned unit and its closest ground-truth phoneme is comparable to the inherent variability of the phoneme. The learned speech units have a good coverage of linguistically defined phonemes. However, there are a few phonemes that cannot be covered, for example, the vowel /er/ in Mandarin, probably due to limited feature representation capacity in the current system. The confusion between ground-truth phonemes can be alleviated with a large cluster number. Further investigation is needed to design new features and segment representation that can better capture trajectory characteristics of phonemes.

7. Acknowledgements

This research is partially supported by a GRF project grant (Ref: CUHK 14227216) from Hong Kong Research Grants Council.

[1] Phoneme label /aa/ is used in OGI-MTS database [31], the corresponding IPA transcription is /a/. Similarly hereinafter.

8. References

[1] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[2] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, "Lattice-based unsupervised acoustic model training," in *Proc. ICASSP*, 2011, pp. 4656–4659.

[3] 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 & Language*, vol. 24, no. 3, pp. 433–444, 2010.

[4] J.-T. Huang and M. Hasegawa-Johnson, "Semi-supervised training of gaussian mixture models by conditional entropy minimization," in *Proc. INTERSPEECH*, 2010, pp. 1353–1356.

[5] J. Glass, "Towards unsupervised speech processing," in *Proc. ISSPA*, 2012, pp. 1–4.

[6] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training." in *Proc. ICASSP*, 2013, pp. 8091–8095.

[7] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proc. ACL*, 2012, pp. 40–49.

[8] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "A graph-based gaussian component clustering approach to unsupervised acoustic modeling." in *Proc. INTERSPEECH*, 2014, pp. 875–879.

[9] S. Feng, T. Lee, and H. Wang, "Exploiting language-mismatched phoneme recognizers for unsupervised acoustic modeling," in *Proc. ISCSLP*, 2016, pp. 1–5.

[10] H. Wang, "Query-by-example spoken term detection for low-resource languages," Ph.D. dissertation, The Chinese University of Hong Kong, Hong Kong, 2014.

[11] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 271–284, 2007.

[12] D. F. Harwath, T. J. Hazen, and J. R. Glass, "Zero resource spoken audio corpus analysis," in *Proc. ICASSP*, 2013, pp. 8555–8559.

[13] M. Versteegh, R. Thiolliere, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015." in *Proc. INTERSPEECH*, 2015, pp. 3169–3173.

[14] Y. Steve, "Large vocabulary continuous speech recognition: a review," *IEEE Signal Processing Magazine*, vol. 21, pp. 786–797, 1996.

[15] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[16] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Proc. INTERSPEECH*, 2014, pp. 338–342.

[17] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in *Proc. ICASSP*, 2008, pp. 3989–3992.

[18] A. H. H. N. Torbati, J. Picone, and M. Sobel, "Speech acoustic unit segmentation using hierarchical dirichlet processes." in *Proc. INTERSPEECH*, 2013, pp. 637–641.

[19] Y. P. Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *Proc. ICASSP*, vol. 4, 2007, pp. 937–940.

[20] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Acoustic segment modeling with spectral clustering methods," *IEEE/ACM Trans. ASLP*, vol. 23, no. 2, pp. 264–277, 2015.

[21] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *Proc. ICASSP*, 1988, pp. 501–504.

[22] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Proc. ICASSP*, 2012, pp. 5157–5160.

[23] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proc. ICASSP*, vol. 2, 1993, pp. 447–450.

[24] G. Aradilla, J. Vepa, and H. Bourlard, "An acoustic model based on Kullback-Leibler divergence for posterior features," in *Proc. ICASSP*, vol. 4, 2007, pp. 657–660.

[25] G. Aradilla, H. Bourlard, and M. M. Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proc. INTERSPEECH*, 2008, pp. 928–931.

[26] T. Asami, R. Masumura, H. Masataki, M. Okamoto, and S. Sakauchi, "Training data selection for acoustic modeling via submodular optimization of joint Kullback-Leibler divergence," in *Proc. INTERSPEECH*, 2015, pp. 3645–3649.

[27] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN approach to cross-lingual TTS," in *Proc. ICASSP*, 2016, pp. 5515–5519.

[28] ——, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. INTERSPEECH*, 2016, pp. 287–291.

[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[30] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[31] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus." in *ICSLP*, vol. 92. Citeseer, 1992, pp. 895–898.

[32] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Brno University of Technology, Brno, Czech Republic, 2009.

[33] T. Lee, W. K. Lo, P. C. Ching, and H. Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, vol. 36, no. 3, pp. 327–342, 2002.

[34] Wikipedia, "Pinyin table — wikipedia, the free encyclopedia," 2017, [Online; accessed 21-March-2017]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pinyin_table&oldid=758841913