Unsupervised Domain Discovery using Latent Dirichlet Allocation for Acoustic Modelling in Speech Recognition

Mortaza Doulaty, Oscar Saz, Thomas Hain

Speech and Hearing Group, University of Sheffield, Sheffield, UK

{mortaza.doulaty, o.saztorralba, t.hain}@sheffield.ac.uk

Abstract

Speech recognition systems are often highly domain dependent, a fact widely reported in the literature. However the concept of domain is complex and not bound to clear criteria. Hence it is often not evident if data should be considered to be out-of-domain. While both acoustic and language models can be domain specific, work in this paper concentrates on acoustic modelling. We present a novel method to perform unsupervised discovery of domains using Latent Dirichlet Allocation (LDA) modelling. Here a set of hidden domains is assumed to exist in the data, whereby each audio segment can be considered to be a weighted mixture of domain properties. The classification of audio segments into domains allows the creation of domain specific acoustic models for automatic speech recognition. Experiments are conducted on a dataset of diverse speech data covering speech from radio and TV broadcasts, telephone conversations, meetings, lectures and read speech, with a joint training set of 60 hours and a test set of 6 hours. Maximum A Posteriori (MAP) adaptation to LDA based domains was shown to yield relative Word Error Rate (WER) improvements of up to 16% relative, compared to pooled training, and up to 10%, compared with models adapted with human-labelled prior domain knowledge.

Index Terms: domain discovery, latent dirichlet allocation, adaptation, speech recognition

1. Introduction

Recently, new applications and domains are becoming the target of research in Automatic Speech Recognition (ASR), as the existing systems increase their accuracy. This has opened the issue on how to scale up existing systems when new domains are incorporated as target data, for instance "found data", such as media and historical audio archives. In this situation, training acoustic models for an unknown domain, like different YouTube recordings, can be infeasible if the origin of the target speech can not be properly assessed, and the loss of accuracy is large due to wrong modelling decisions.

Well–tailored single domain systems, where training data that properly matches the target recognition data is available, are mostly used in current speech recognisers. These domain dependent models have been usually trained via Maximum Likelihood (ML) if a sufficiently large amount of domain data existed or using adaptation techniques such as Maximum A Posteriori [1], Maximum Likelihood Linear Regression (MLLR) [2] or Cluster Adaptive Training [3]. For more recent Deep Neural Network (DNN)–based systems, domain adaptation is also possible with linear transformations, conservative training and subspace methods [4] with frameworks such as Multi–Level Adaptive Networks (MLAN) [5] or Deep Maxout Networks (DMN) [6].

An important issue when dealing with highly diverse speech data is the difficulty to appropriately categorise every speech input within a particular domain, especially the case with newly discovered data. Even when domain categories have been given manually by humans, this may be inaccurate or there may be hidden characteristics in the audio that can further subdivide these categories or cross across several of the predefined domains. Developing the ability of discovering these new and hidden acoustic domains would greatly enhance the possibility of using well–targeted specific domain models in ASR. However, as most speech recognition tasks assume a single domain or well differentiated domains, the task of unsupervised discovery of acoustic domains in speech data has been of less interest so far. This paper proposes to open new areas for research in multi–domain ASR by treating speech data as a set of documents where latent domains exist and can be discovered using Latent Dirichlet Allocation (LDA) models.

LDA is an statistical approach to discover latent topics in a collection of documents in an unsupervised manner [7]. It is mostly used in Natural Language Processing (NLP) for the categorisation of text documents, but it has been used for audio and image processing as well. In audio tasks, LDA has been used for classifying unstructured audio files into onomatopoeic and semantic descriptions with successful results [8, 9]. Building on this knowledge, this work proposes to use LDA for domain adaptation in ASR tasks.

This paper is organised as follows: Section 2 will give an overview of LDA modelling in its original proposal for topic modelling. Then, Section 3 will describe the proposed use of LDA models for unsupervised domain discovery in speech data. Section 4 will present the experimental setup used for multi–domain speech recognition, with Section 5 detailing the obtained results. Section 6 gives the conclusions to this work.

2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [7] is an unsupervised probabilistic generative model for collections of discrete data. It aims to describe how every item within the collection is generated, assuming that there are a set of hidden topics and that each item is modelled as a finite mixture over those topics. Also, an infinite mixture over an underlying set of topic probabilities is used to model each topic [7]. LDA is mostly used for topic modelling of text corpora, however, the model can be applied to other tasks, such as object categorisation and localisation in image processing [10], automatic harmonic analysis in music processing [11] or acoustic information retrieval in unstructured audio analysis [9].

In the context of text corpora, a dataset is defined as a collection of documents and each document is a collection of words. Given a vocabulary of size V, each word is represented by a V–dimensional binary vector. It is assumed that the docu-

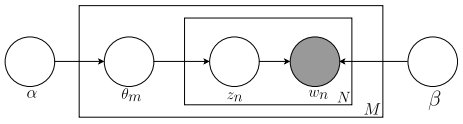

Figure 1: Graphical model representation of LDA

ments are generated using the following generative process:

1. For each document $d_m, m \in \{1...M\}$, choose a K–dimensional topic weight vector θ_m from the Dirichlet distribution with scaling parameter α: $p(\theta_m|\alpha) = Dir(\alpha)$
2. For each word $w_n, n \in \{1...N\}$ in document d_m
   (a) Draw a topic $z_n \in \{1...K\}$ from the multinomial distribution $p(z_n = k|\theta_m)$
   (b) Given the topic, draw a word from $p(w_n|z_n, \beta)$, where β is a $V \times K$ matrix and $\beta_{ij} = p(w_n = i|z_n = j, \beta)$

Other assumptions include the bag–of–words property of the documents and the fixed and known dimensionality of the Dirichlet distribution K (and thus the dimensionality of the topic variable z)

The graphical representation of LDA model is shown at Figure 1, a three level hierarchical Bayesian model. In this model, the only observed variable is w and the rest are all latent. α and β are corpus level parameters, θ_m are document level variables and z_n, w_n are word level variables. The generative process is described formally as:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \quad (1)$$

The posterior distribution of the latent topic variables given the words and α and β parameters is:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \quad (2)$$

Computing $p(\mathbf{w}|\alpha, \beta)$ requires some intractable integrals. A reasonable approximate can be acquired using variational approximation which is shown to work reasonably well in various applications [7]. The approximated posterior distribution is:

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \quad (3)$$

where γ is the Dirichlet parameter that determine θ and ϕ is the parameter for the multinomial that generates the topics.

Training tries to minimise the Kullback–Leiber divergence (KLD) [12] between the real and the approximated joint probabilities (equations 2 and 3) [7]:

$$\underset{\gamma,\phi}{argmin}\ KLD\big(q(\theta, \mathbf{z}|\gamma, \phi)\ ||\ p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)\big) \quad (4)$$

Other training methods based on Markov–Chain Monte-Carlo is also proposed, like Gibbs sampling method [13].

3. Unsupervised Domain Discovery

The proposed technique uses an LDA model to discover hidden and latent acoustic domains in multi–domain speech data. Since LDA is for collections of discrete data (such as text corpora) [7], every speech segment of length T frames, $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_T\}$, is represented as a set of discrete symbols to support modelling within this framework. For that purpose, the n–dimensional audio frames, $\mathbf{x}_t \in \mathbb{R}^n$, are quantised into a dictionary of V acoustic "words", $\bar{\mathbf{x}}_t \in \{1...V\}$ [8]. First a Gaussian Mixture Model (GMM) is trained using Expectation Maximisation (EM) and mix–up procedure to reach the desired codebook size V (enforcing the co–variance matrix to be identity, equivalent to LBG–VQ [14]). Then the means of the Gaussian components are used to create the codebook and quantise the audio frames into discrete symbols. The assignment of frame $\mathbf{x}_i$ to codebook index j is performed using:

$$\bar{\mathbf{x}}_t = \underset{j}{argmin}\ ||\mathbf{x}_t - \mathbf{m}_j||\ , j \in \{1...V\} \quad (5)$$

where $\mathbf{m_j}$ is jth mixture component's mean vector.

To reconcile this with the LDA terminology described in Section 2, in this work each audio segment is a "document" and each codified audio frame is a "word". All the audio segments (now "documents") then create a whole "collection" or "corpus".

Once all the audio frames are converted to discrete "words", the parameters of the LDA model using K domains are estimated on the M audio segments from the training data using variational EM. The domain of each quantised audio segment $\bar{\mathbf{x}}$ is then given by the domain with the highest value of the posterior Dirichlet parameter γ for that segment.

$$Domain(\bar{\mathbf{x}}) = \underset{j}{argmax}\ \gamma_j\ ,\ j \in \{1..K\} \quad (6)$$

Based on the estimated parameters from the training set, Dirichlet parameters γ can be inferred for the test set segments as well. With every segment in both train and test sets associated to a hidden domain, it is possible to perform training and/or adaptation with the usual techniques. Acoustic models can be trained via Maximum Likelihood (ML), or domain specific models can be adapted via MAP or MLLR, in case of GMM/HMM systems.

4. Experimental setup

To evaluate the proposed domain discovery and adaptation method in a multi–domain and diverse ASR task, a dataset of 6 different types of data was chosen from the following sources:

- Radio (RD): BBC Radio4 broadcasts on February 2009.
- Television (TV): Broadcasts from BBC on May 2008.
- Telephone speech (CT): From the Fisher corpus[1] [15].
- Meetings (MT): From AMI [16] and ICSI [17] corpora.
- Lectures (TK): From TedTalks [18].
- Read speech (RS): From the WSJCAM0 corpus [19].

A subset of 10h from each domain was selected to form the training set (60h in total), and 1h from each domain was used for testing (6h in total). The selection of the domains aims to cover the most common and distinctive types of audio recordings used in ASR tasks.

Two types of acoustic features were used: First, 13 PLP features plus first and second derivatives for a total of 39–dimensional feature vectors; and second, a 65–dimensional feature vector concatenating the 39 PLP features and 26 bottleneck (PLP+BN) features extracted from a 4–hidden–layer DNN trained on the full 60 hours of data. 31 adjacent frames (15

[1] All of the telephone speech data was up–sampled to 16 kHz to match the sampling rate of the rest of the data.

frames to the left and 15 frames to the right) of 23 dimensional log Mel filter bank features were concatenated to form a 713–dimensional super vector; Discrete Cosine Transform (DCT) was applied to this super vector to de–correlate and compress it to 368 dimensions and then fed into the neural network. The network was trained on 4,000 triphone state targets and the 26 dimensional bottleneck layer was placed before the output layer. The objective function used was frame–level cross–entropy and the optimisation was done with stochastic gradient descent and the backpropagation algorithm. DNN training was performed with the TNet toolkit [20] and more details can be found at [21].

For both types of features, baseline ML GMM–HMM models were trained using HTK [22] with 5–state crossword triphones and 16 gaussians per state. The language model used was based on a 50,000–word vocabulary and was trained by combination of language models from the 6 domains, with interpolation weights tuned using an independent development set.

4.1. Baseline results

Table 1 presents the baseline Word Error Rate (WER) results for the in–domain maximum–likelihood (ML) model trained with the pooled 60 hours of all domains, plus the results of ML in–domain models each trained with 10 hours of in–domain data. It also includes the MAP adapted models from the pooled model to each domain. Experiments were conducted using PLP and PLP+BN features. The results using ML training on the limited in–domain data underperformed MAP adaptation on such data, which set MAP as a preferred setup for domain adaptation.

Table 1: WER (%) of baseline models

Features	Model	RS	RD	TK	CT	MT	TV	Total
PLP	ML	17.3	18.4	34.1	46.6	44.0	51.1	**36.0**
	ML Domain	16.9	19.1	35.1	44.4	44.0	52.9	**36.3**
	MAP	14.6	16.8	31.8	43.5	40.4	49.6	**33.6**
PLP+BN	ML	13.0	13.3	23.5	33.5	32.2	42.0	**26.8**
	ML Domain	12.6	14.0	25.0	34.3	33.2	44.0	**27.9**
	MAP	12.1	12.8	23.1	32.5	30.6	41.5	**26.2**

5. Results

The experiments performed aimed to evaluate two aspects of the proposed LDA modelling for unsupervised domain discovery. First, if LDA could be successfully used to find hidden domains and if these domains represented the hidden characteristics of the audio. Second, once hidden domains had been identified, if domain adaptation could be applied on them and improvements in ASR performance were achieved over the baselines.

5.1. Unsupervised domain discovery

For using LDA models, as described in Section 2, two parameters had to be initially set up. First, the number of domains K to be found had to be decided prior to the training. Also, since the audio frames needed to be quantised, the size of the codebook V also needed to be defined. For this end, a set of experiments were conducted with different codebook sizes and number of domains. Codebooks of size 128 up to 8,192 were used and given a codebook, different LDA models with a varying number of domains from 4 to 64 were estimated [23, 24] using the training data described in Section 4.

Since these identified domains were latent, there was no ground truth to verify them at this stage. An initial way of evaluating how the different latent domains behaved was by measuring the distribution of the data, according to manual labels,

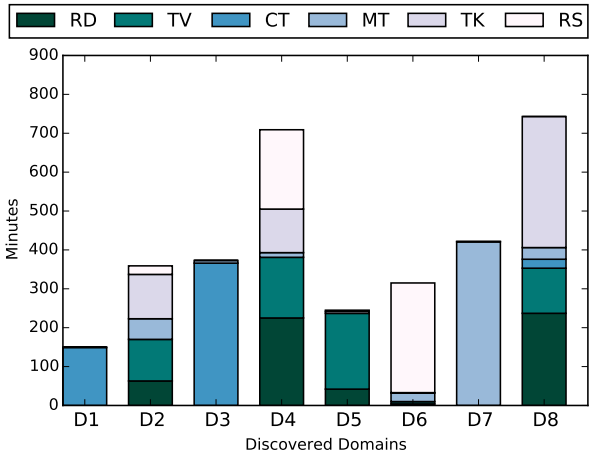

Figure 2: Amount of data for each discovered domain ($K = 8$) from the labelled domains using a codebook size of 2,048

which was included in each hidden domain. Figure 2 presents this distribution for an acoustic codebook of size 2,048 and 8 hidden domains. From this Figure, it is possible to see how telephone speech was separated into two different hidden domains (D1 and D3), while meeting speech was mostly assigned to a unique hidden domain (D7). Other manually labelled domains, such as Radio and Television broadcasts were scattered across hidden domains (D2, D4 or D8), indicating the presence of previously unseen domains within these types of data.

Following this, KL divergence [12] was proposed as an appropriate metric to measure the consistency of the hidden topics discovered by LDA. This measured how the distributions of data in latent domains, as in Figure 2, in different sets, for instance training and testing data, were different with each other:

$$KLD(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \quad (7)$$

where P and Q are the distributions for training and test data. To compute the divergence, since we deal with counts in the distributions and some counts can be zero, the distributions are smoothed by discounting 3% of the total mass and distributing it across zero counts.

Figure 3 shows the divergence values of different configurations. Low values of divergence indicated a more consistent set of hidden domains found by LDA modelling and, thus, were preferred over configurations with higher values. In terms of codebook size, codebooks of 2,048 and 8,192 symbols resulted in lower divergence. For the number of domains, increasing to more than 12 resulted an increase in divergence.

5.2. Domain adaptation

For the evaluation of the possibilities offered by the unsupervised discovery of domains in ASR, MAP domain adaptation was performed to each of these new domains. The experiments were conducted with domains of size 4, 6, 8, 10 and 12 and a codebook of acoustic words of size 2,048. Each MAP adapted domain specific model was used to decode the corresponding speech segments in the test set that were assigned to that domain. Figure 4 shows the overall WER on the test set with different number of topics using both types of features, PLP and PLP+BN. The lowest WER values, 30.4% for PLP features and

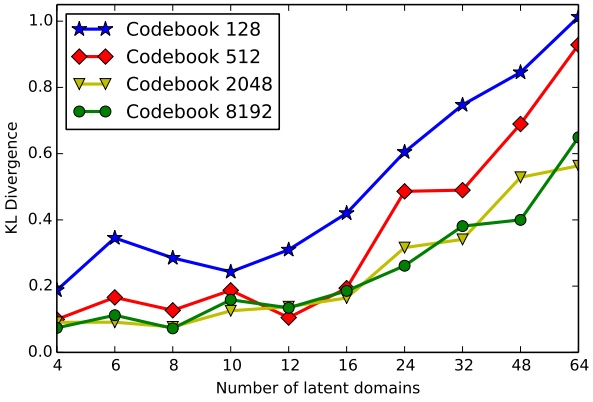

Figure 3: KL divergence of training and test set topics

25.4% for PLP+BN, were achieved with 8 domains for both types of features, which was 16% and 5% relative improvement over their respective ML baselines. Comparing with MAP adaptation to human–labelled domains the relative WER reduction was 10% and 3%. The improvements in WER vanished for more than 8 hidden domains, indicating that using larger numbers of domains were not beneficial for this task.

Table 2 presents the breakout of the results using 8 hidden domains across the manually labelled domains. Improvements occur across all of these domains, indicating that the LDA model can benefit all types of speech in this setup. The domains that achieved the highest gains from using LDA MAP adaptation (with PLP feature) were read speech, telephone speech and TV broadcasts, with relative WER reductions of 14%, 12%, 10% respectively compared to MAP adaptation on the manually labelled domains. The lowest gain, 4% relative, occurred on meeting speech. Similarly, with PLP+BN features telephone speech, lectures and read speech benefited the most, with relative WER reduction of 5%, 4% and 2% respectively.

Table 2: WER (%) of LDA MAP Models ($K = 8$)

Features	Model	RS	RD	TK	CT	MT	TV	Total
PLP	MAP	14.6	16.8	31.8	43.5	40.4	49.6	**33.6**
	LDA MAP	12.5	15.3	29.1	38.2	38.5	44.7	**30.4**
PLP+BN	MAP	12.1	12.8	23.1	32.5	30.6	41.5	**26.2**
	LDA MAP	11.9	12.8	22.3	31.1	31.0	41.0	**25.4**

Finally, Table 3 shows the WER across the hidden domains for both types of features with LDA MAP models. The most relevant feature of these domains, in terms of WER, was that the domains of low WER (like Read speech) or high WER (like TV data) had been broken up in different hidden domains and hence, WERs across hidden domains were evenly distributed.

Table 3: WER (%) of LDA MAP Models ($K = 8$) across hidden domains

Features	D1	D2	D3	D4	D5	D6	D7	D8	Total
PLP	37.3	34.9	39.7	39.2	24.6	17.1	38.7	22.9	**30.4**
PLP+BN	33.9	29.2	30.4	32.8	19.7	12.6	30.9	19.2	**25.4**

6. Conclusions

A novel technique based on Latent Dirichlet Allocation (LDA) has been proposed to discover latent domains in highly–diverse speech data in an un–supervised manner. The data set consisted

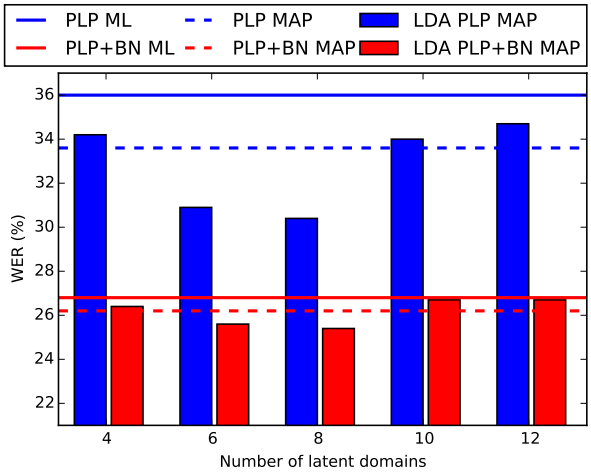

Figure 4: WER (%) of LDA MAP adapted models with different number of topics

of data from TV and radio shows, meetings, lectures, talks and telephony speech with a 60–hour training set and 6–hour test set. It was assumed that there are a set of hidden domains and each audio segment is a mixture of different properties of those hidden domains with different weights. LDA models were used to discover the latent domains and then these domains were used to perform Maximum A Posteriori (MAP) domain adaptation. Results showed relative improvement of up to 16% over the baseline Maximum Likelihood trained models and up to 10% over the MAP adapted models to human labelled domains with the LDA discovered domains.

The bag–of–words assumption in LDA model does not take the order of words into account. In applying LDA for image processing, there are some variants of the original LDA model, such as Spatial LDA [25] which encodes spatial structure with the visual words. A temporal variant of LDA could better handle the temporal nature of speech and needs to be investigated as a future work. Also applying the current technique on bigger and/or less diverse data set needs to be verified to see what would be the new discovered domains and how they are related to domain adaptation. Newer sets of features, better targeted to describe background acoustic characteristics [26], could also provide an improvement over PLP features, which are known to describe well phonetic and speaker information.

7. Acknowledgements

This work was supported by the EPSRC Programme Grant EP/I031022/1 Natural Speech Technology (NST).

8. Data Access Statement

The speech data used in this paper was obtained from the following sources: Fisher Corpus (LDC catalogue number LDC2004T19), ICSI Meetings corpus (LDC2004S02), WSJCAM0 (LDC95S24), AMI corpus (DOI number 10.1007/11677482_3), TedTalks data (freely available as part of the IWSLT evaluations), BBC Radio and TV data (this data was distributed to the NST project's partners with an agreement with BBC R&D and not publicly available yet).

The specific file lists used for training and testing in the experiments in this paper, as well as result files can be downloaded from `http://mini.dcs.shef.ac.uk/publications/papers/is15-doulaty2`.

9. References

[1] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[3] M. J. Gales, "Cluster adaptive training for speech recognition." in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP–Interspeech)*, Sydney, Australia, 1998, pp. 1783–1786.

[4] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. London, UK: Springer-Verlag, 2015.

[5] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Proceddings of the 2013 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver BC, Canada, 2013, pp. 6975–6979.

[6] Y. Miao, F. Metze, and S. Rawat, "Deep maxout networks for low-resource speech recognition," in *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp. 398–403.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[8] S. Kim, S. Sundaram, P. Georgiou, and S. Narayanan, "Audio scene understanding using topic models," in *Proceedings of the Neural Information Processing System (NIPS) Workshop*, Whistler BC, Canada, 2009.

[9] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic model for audio information retrieval," in *Proceedings of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, 2009, pp. 37–40.

[10] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proceedings of the 10th International Conference on Computer Vision (ICCV)*, Beijing, China, 2005, pp. 370–377.

[11] D. Hu and L. K. Saul, "A probabilistic topic model for unsupervised learning of musical key-profiles." in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 2009, pp. 441–446.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 2, no. 1, pp. 79–86, 1951.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 5228–5235, 2004.

[14] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Berlin, Germany: Springer Science & Business Media, 1992.

[15] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text." in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004, pp. 69–71.

[16] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, W. Karaiskos, Vasilis Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proceedings of the Third International Workshop on Machine Learning for Multimodal Interaction*, Bethesda MD, USA, 2006, pp. 28–39.

[17] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proceedings of the 2003 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 2003.

[18] R. W. N. Ng, M. Doulaty, R. Doddipatla, O. Saz, M. Hasan, T. Hain, W. Aziz, K. Shaf, and L. Specia, "The USFD spoken language translation system for IWSLT 2014," Lake Tahoe NV, USA, 2014.

[19] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A british english speech corpus for large vocabulary continuous speech recognition," in *Proceedings of the 1995 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit MI, USA, 1995.

[20] K. Vesely, L. Burget, and F. Grezl, "Parallel training of neural networks for speech recognition," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, Makuhari, Japan, 2010.

[21] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proceedings of the 2014 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, *The HTK book (for HTK version 3.4)*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[23] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *Proceedings of the Neural Information Processing System (NIPS) Workshop*, Vancouver BC, Canada, 2010, pp. 856–864.

[24] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, 2010, pp. 45–50.

[25] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *Proceedings of the Neural Information Processing System (NIPS) Workshop*, Whistler BC, Canada, 2008, pp. 1577–1584.

[26] O. Saz, M. Doulaty, and T. Hain, "Background–tracking acoustic features for genre identification of broadcast shows," in *Proceedings of the 2014 IEEE Workshop on Spoken Language Technologies (SLT)*, Lake Tahoe NV, USA, 2014.