Augmented Conditional Random Fields Modeling Based on Discriminatively Trained Features

Yasser Hifny

University of Helwan, Egypt
yhifny@fci.helwan.edu.eg

Abstract

Augmented Conditional Random Fields (ACRFs) are undirected graphical models that maintain the Markov properties of Hidden Markov Models (HMMs), formulated using the maximum entropy (MaxEnt) principle. ACRFs incorporate acoustic context information into an augmented space in order to model the sequential phenomena of the speech signal. The augmented space is constructed using Gaussian activation functions representing the dense regions in the observation space. These activation functions are estimated using the Expectation-Maximization (EM) algorithm. Alternatively, the activation functions can be estimated using a discriminative objective function. Hence, the ACRFs are fed with discriminative features. In this paper, we show that ACRFs recognition results improve if the activation functions are estimated using the Minimum Phone Error (MPE) discriminative criterion.

Index Terms: Augmented Conditional Random Fields (ACRFs), discriminative training, augmented features

1. Introduction

Over the last few years, there is an increased interest to develop acoustic models derived from Conditional Random Fields [1]. In [2, 3], a new acoustic modeling paradigm based on Augmented Conditional Random Fields (ACRFs) is investigated and developed. Other alternatives were developed in [4],[5],[6], [7], [8]. Several MaxEnt/CRF methods were reviewed and discussed in [9]. ACRFs paradigm addresses some limitations of HMMs while maintaining many of the aspects which have made them successful. In particular, the acoustic modeling problem is reformulated in a data driven, sparse, augmented space to increase discrimination. Acoustic context modeling is explicitly integrated to handle the sequential phenomena of the speech signal. On the other hand, it is possible to have fast and efficient discriminative training for ACRFs.

ACRFs incorporate acoustic context information into an augmented space in order to model the sequential phenomena of the speech signal. The augmented space is constructed using Gaussian activation functions representing the dense regions in the observation space. These activation functions are estimated using the EM algorithm. Alternatively, the activation functions can be estimated using a discriminative objective function. Hence, the ACRFs are fed with discriminative features.

In Section 2, a mathematical formulation of ACRFs is described. The optimization problem of ACRFs is addressed in Section 3. Feature projection into high dimensional spaces (i.e Augmented Spaces) is presented in Section 4. Section 5 gives experimental results on a phone recognition task. Finally, a summary of the presented work is given in the conclusions.

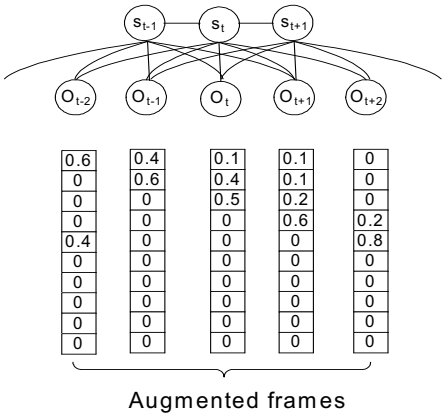

Figure 1: ACRF model for phone representation that is dependent on arbitrary acoustic observations. Modeling the acoustic context by concatenating a window of multiple acoustic frames is equivalent to having a state dependence on the previous/following c frames ($c = 2$ in this ACRF model).

2. Augmented Conditional Random Fields

Augmented Conditional Random Fields (ACRFs) incorporate acoustic context information into an augmented space in order to model the sequential phenomena of the speech signal. ACRFs are derived from linear chain CRFs[1] [1]. Linear chain CRFs can be thought as the undirected graphical twins for HMMs regardless of their training (generative or discriminative). ACRF acoustic models are a particular implementation of linear chain CRFs developed for augmented acoustic spaces. A graphical representation of the ACRF acoustic model is shown in Figure 1; its states are dependent on arbitrary acoustic observations. The conditional independence properties of the HMM are relaxed explicitly in the ACRF acoustic model with the sufficient statistics collected from an augmented space.

The conditional distribution defining ACRFs is given by

$$P_\Lambda(\mathbf{S}|\mathbf{O}) = \frac{1}{Z_\Lambda(\mathbf{O})} \prod_{t=1}^{T} \exp\left(\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}} a(\mathbf{s}_t, \mathbf{s}_{t-1}) + \sum_{u=t-c}^{t+c} \sum_{i=1}^{d^{\text{Aug}}} \lambda_{\mathbf{s}_t}^{ui} b_i(\mathbf{o}_u, \mathbf{s}_t)\right) \quad (1)$$

where

[1] More precisely, ACRFs are a nonlinear form of linear chain CRFs.

- $P_\Lambda(\mathbf{S}|\mathbf{O})$ obeys the Markovian property:

$$P_\Lambda(\mathbf{s}_t|\{\mathbf{s}_j\}_{j\neq t},\mathbf{O}) = P_\Lambda(\mathbf{s}_t|\mathbf{s}_{t-1},\mathbf{O})$$

- $\lambda^{ui}_{\mathbf{s}_t}$ and $\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}}$ are associated with the characterizing functions $b_i(\mathbf{o}_u,\mathbf{s}_t)$ and $a(\mathbf{s}_t,\mathbf{s}_{t-1})$.

- $w = 2c+1$ is the number of frames in the acoustic context window

- $Z_\Lambda(\mathbf{O})$ (Zustandsumme) is a normalization coefficient referred to as the partition functions.

- $b_i(\mathbf{o}_t,\mathbf{s}_t)$ is computed in Equation (8).

HMMs and ACRFs (in general, linear chain CRFs) share the first order Markov assumption, which simplifies the training and decoding algorithms. However, unlike HMMs, ACRFs do not assume observation independence and causality, as the joint event in this case is factorized as a simple product of exponential functions. Therefore, the observations and the characterizing functions can be statistically dependent or correlated and can depend on the past and future acoustic context. As a result, ACRFs provide a principled way to relax the HMM conditional independence assumption. The partition function, $Z_\Lambda(\mathbf{O})$, is given by

$$Z_\Lambda(\mathbf{O}) = \sum_{\mathbf{S}}\prod_{t=1}^{T} \exp\Big(\lambda_{\mathbf{s}_t\mathbf{s}_{t-1}}a(\mathbf{s}_t,\mathbf{s}_{t-1}) + \sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}}\lambda^{ui}_{\mathbf{s}_t}b_i(\mathbf{o}_u,\mathbf{s}_t)\Big), \quad (2)$$

and it is similar to the total probability $p(\mathbf{O}|\mathcal{M})$ in HMMs, which can be calculated using the forward algorithm [1].

The ACRF model takes advantage of the construction of augmented spaces to model the acoustic context. It may be expected that modeling acoustic context in augmented spaces within the ACRF framework is an effective technique since the augmented space confusability is expected to be less than for low dimensional spaces. This additional context may increase discrimination within the acoustic modeling process.

3. ACRF Optimization

For R training observations $\{\mathbf{O}_1,\mathbf{O}_2,\ldots,\mathbf{O}_r,\ldots,\mathbf{O}_R\}$ with corresponding transcriptions $\{W_r\}$, ACRFs are trained using the conditional maximum likelihood (CML) criterion to maximize the posterior probability of the correct word sequence given the acoustic observations. In our work, we use a family of iterative scaling algorithms, which we call *Approximate Iterative Scaling* (AIS), to speed up the training process. An AIS algorithm updates an e-family activation function based on first-order sufficient statistics, as shown in the following equation:

$$\lambda^{\tau+1}_{ji}(\mathbf{O}) = \lambda^{\tau}_{ji}(\mathbf{O}) + \eta_{\text{AIS}}\log\frac{\mathcal{C}^{\text{num}}_{ji}(\mathbf{O})}{\mathcal{C}^{\text{den}}_{ji}(\mathbf{O})} \quad (3)$$

where $\eta_{\text{AIS}} = \frac{1}{w}$ is called the *learning rate* and τ is the iteration number. The sparse accumulators of the sufficient statistics, $\mathcal{C}_{ji}(\mathbf{O})$, for the j^{th} state and i^{th} constraint are calculated as follows:

$$\mathcal{C}^{\text{num}}_{ji}(\mathbf{O}) = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma^r_j(t|\mathcal{M}^{\text{num}})\mathbf{o}^{\text{Aug}}_{rti} \quad (4)$$

$$\mathcal{C}^{\text{den}}_{ji}(\mathbf{O}) = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma^r_j(t|\mathcal{M}^{\text{den}})\mathbf{o}^{\text{Aug}}_{rti} \quad (5)$$

where r is the utterance index and the frame-state alignment probability γ_j, denoting the probability of being in state j at some time t can be written in terms of the forward score $\alpha_j(t)$ and the backward score $\beta_j(t)$ as in HMMs:

$$\gamma_j(t|\mathcal{M}) = P(\mathbf{s}_t = j|\mathbf{O};\mathcal{M}) = \frac{\alpha_j(t|\mathcal{M})\beta_j(t|\mathcal{M})}{Z_\Lambda(\mathbf{O}|\mathcal{M})} \quad (6)$$

and to avoid the necessity of building lattices, the $\gamma_j(t|\mathcal{M})$ is approximated with state estimates as follows [10]:

$$\gamma_j(t|\mathcal{M}) = \frac{\exp\big(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}}\lambda^{ui}_j b_i(\mathbf{o}_u,j)\big)}{\sum_{\mathbf{s}}\exp\big(\sum_{u=t-c}^{t+c}\sum_{i=1}^{d^{\text{Aug}}}\lambda^{ui}_{\mathbf{s}} b_i(\mathbf{o}_u,\mathbf{s})\big)} \quad (7)$$

4. Augmented Spaces

4.1. Generatively trained activation functions

Feature projection into high dimensional spaces is a powerful tool to simplify classification problems, since high dimensional spaces are more likely to be linearly separable than low dimensional spaces [11]. The process of augmenting the low dimensional space to result in a high dimensional space $\mathbf{o}_t \rightarrow \mathbf{o}^{\text{Aug}}_t$, starts with the application of a large number of constraints $g_i(\mathbf{o}_t;\lambda)$ to the observed data. Then, the constraints are sorted according to their scores and only the n-best are retained.[2] The indices of the top n constraints represent the kernel functions that are most responsible for activating the acoustic frame under consideration (i.e. the acoustic regions most likely to account for the current frame). The selection of an n-best shortlist is essential to reduce the storage requirements of the approach. Once such a shortlist is available, the augmented vector is constructed and its size d^{Aug} equals the number of constraints in the recognition problem. A state constraint value in the new augmented space is calculated as a pruned posterior score for each parametric constraint and is given by

$$b_i(\mathbf{o}_t,\mathbf{s}_t) = \frac{g_i(\mathbf{o}_t;\lambda)}{\sum_j g_j(\mathbf{o}_t;\lambda)} \approx \frac{g_i(\mathbf{o}_t;\lambda)}{\sum_{j\in n-\text{best}} g_j(\mathbf{o}_t;\lambda)} \quad (8)$$

where the normalization step is conceptually redundant to improve the ACRFs training speed. We use diagonal Gaussian density functions, estimated using the EM algorithm to locate the dense regions:

$$g_i(\mathbf{o}_t;\lambda) = p_i(\mathbf{o}_t \mid \theta) = \mathcal{N}(\mathbf{o}_t;\mu_i,\Sigma_i) \quad (9)$$

The augmentation process $\mathbf{o}_t \rightarrow \mathbf{o}^{\text{Aug}}_t$ is sketched in Figure 2. The dense regions, those regions where most data points are projected, are defined using the hyperellipsoids derived from the eigen decomposition of the covariance matrix in case of the Gaussian activation. In addition, the orientation of the hyperellipsoid axes associated with diagonal covariance Gaussians are parallel to the coordinate axes. The pruned posterior scores are obtained for Gaussians near the point X. Consequently, most of the elements of an augmented vector $\mathbf{o}^{\text{Aug}}_t$ are zero as they are considered outliers for the point X. Moreover, the sum of the elements in an augmented vector is equal to 1.0.

[2] Typically, the n-best nearest-neighbor shortlist size is set to 10.

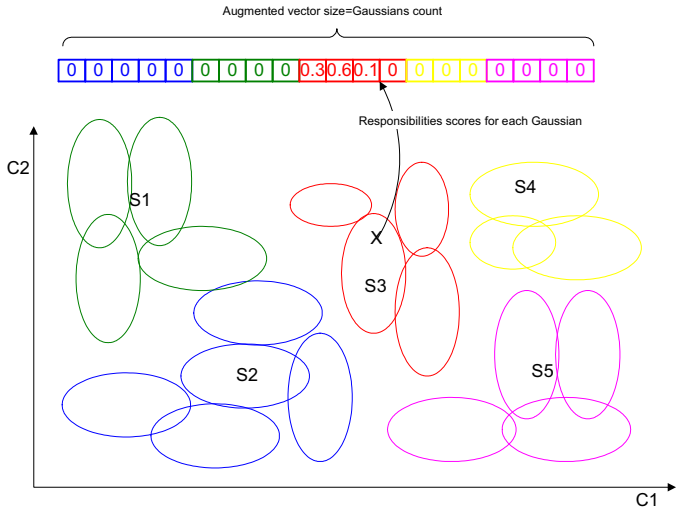

Figure 2: A two dimensional space is partitioned into 20 regions specified by diagonal Gaussians. The augmented space dimensionality is equal to the number of Gaussians (constraint functions) in the acoustic space. Hence, the two dimension space is augmented to a twenty dimensional space $\mathbb{R}^2 \rightarrow \mathbb{R}^{20}$. The augmented vector is constructed by calculating a posterior score for each Gaussian. The majority of the elements of the augmented vector have very low posterior score (i.e. $\approx$ zero). In this case, the n-best shortlist size is set to 3.

4.2. Discriminatively trained activation functions

The diagonal Gaussians can be estimated efficiently using the EM algorithm for GMM/HMM models - based on the maximum likelihood (MLE) criterion- to locate the dense regions in the acoustic space [12]. This is due to the fact that GMMs and continuous density HMMs may be interpreted as soft vector quantizers and the estimated Gaussians represent the dense regions in the clustering process. As a result, by ignoring all mixture weights or transition matrices in an acoustic model, we have a practical method to locate the dense regions in any acoustic space. This is the experimental approach taken in the ACRFs basic formulation presented in [2, 3].

Discriminative training can be an alternative method to estimate the diagonal Gaussian activation function. The Conditional Maximum Likelihood (CML) criterion, defined by equation (10), aims to maximize the log of posterior probability of the correct word sequence given the observations,

$$\begin{aligned}\mathcal{F}_{\mathrm{CML}}(\Lambda) &= \sum_{r=1}^{R} \log P_\Lambda(\mathcal{M}_{w_r}|\mathbf{O}_r) \\ &= \sum_{r=1}^{R} \log \frac{p_\Lambda(\mathbf{O}_r|\mathcal{M}_{w_r})P(w_r)}{\sum_{\hat{w}} p_\Lambda(\mathbf{O}_r|\mathcal{M}_{\hat{w}})P(\hat{w})} \\ &\approx \sum_{r=1}^{R} \log p_\Lambda(\mathbf{O}_r|\mathcal{M}_r^{\mathrm{num}}) - \log p_\Lambda(\mathbf{O}_r|\mathcal{M}_r^{\mathrm{den}})\end{aligned} \tag{10}$$

where $\mathcal{M}_w$ is a composite model corresponding to the word sequence w and $P(w)$ is the probability of this sequence as determined by a language model. This discriminative training aims to maximize a term related to the probability of the correct models (known as the numerator) $p_\Lambda(\mathbf{O}_r|\mathcal{M}^{\mathrm{num}})$, which is identical to the ML objective function, and simultaneously minimize a term related to all incorrect models probabilities (known as the denominator term) $p_\Lambda(\mathbf{O}_r|\mathcal{M}^{\mathrm{den}}) \approx \sum_{\hat{w}} p_\Lambda(\mathbf{O}_r|\mathcal{M}_{\hat{w}})P(\hat{w})$.

Discriminative training based on the CML objective function does not directly measure the expected WER criterion. Instead, the Overall Risk Criterion Estimation (ORCE) [13, 14] directly minimizes the expected word or phone error rates by refining the model parameters based on a measure of risk related to recognition error. Minimum Phone Error (MPE) criterion may be considered as a particular realization of ORCE and it is given by

$$\mathcal{F}_{\mathrm{MPE}}(\Lambda) = \sum_{r=1}^{R} \sum_{w} P_\Lambda(\mathcal{M}_w|\mathbf{O}_r)A(w, w_r) \tag{11}$$

where $A(w, w_r)$ is the raw phone transcription accuracy of the sentence w, given the reference sentence w_r. It has been reported that ORCE based on MPE criterion gives a small improvement over ORCE based on Minimum Word Error (MWE) criterion [15, 16].

In this work, we train HMMs using MPE criterion [16]. We ignore all mixture weights or transition matrices in an acoustic model and use the resultant Gaussians to construct the augmented spaces.

Hence, we have two methods to construct the augmented spaces; one is based on the MLE criterion and the other method is based on the MPE criterion. The two methods will be investigated in the following section.

5. Experiments

We have carried out phone recognition experiments on the TIMIT corpus.[3] We used the 462 speaker training set and testing on the 24 speaker core test set (the SA1 and SA2 utterances were not used). The speech was analyzed using a 25ms Hamming window with a 10 ms fixed frame rate. In all the experiments we represented the speech using 12th order mel frequency cepstral coefficients (MFCCs), energy, along with their first and second temporal derivatives, resulting in a 39 element feature vector. Following Lee [17], the original 61 phone classes in TIMIT were mapped to a set of 48 labels, which were used for training. This set of 48 phone classes was mapped down to a set of 39 classes [17], after decoding, and phone recognition results are reported on these classes, in terms of the phone error rate (PER), which is analogous to word error rate. All our experiments used a bigram language model over phones, estimated from the training set. The language model scaling factor is set to 6.0 during the decoding process.

The baseline HMMs have three emitting states and the emission probabilities were modeled with mixtures of Gaussian densities with diagonal covariance matrices. The generative HMMs were trained by the maximum likelihood criterion using the conventional EM algorithm [18]. HMMs were refined using MPE discriminative training based on a variant of the EBW algorithm [16]. The acoustic scale was set to $1/6$ and τ was set to 100.

The Gaussian set in the HMM based acoustic models are used to construct augmented features for ACRFs modeling as described in section 4. Hence, ACRFs results are reported for the Gaussian set trained using MLE criterion and the Gaussian set trained using MPE criterion. Each phone was represented

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

Table 1: Discriminatively trained activation functions performance on TIMIT recognition task in terms of PER.

Complexity	HMM		ACRF	
#Gaussian	MLE	MPE	MLE	MPE
288	41.0%	39.9%	32.5%	31.9%
1440	33.4%	32.9%	28.6%	28.6%
2880	31.6%	31.0%	28.7%	28.4%
5760	30.4%	30.3%	27.9%	28.3%

using a three state left-to-right ACRF, all parameters related to the constraint functions were initialized to zero and the transition probabilities were initialized either from trained HMM models or from a uniform transition matrix, forcing left to right ACRFs. The training procedure accumulated the $\mathcal{M}^{\text{num}}$ sufficient statistics via a Viterbi pass (forced alignment) of the reference transcription using HMMs trained using maximum likelihood criterion. 20 iterations were used to train ACRFs and the language model scaling factor is set to 1.0 during the decoding process. The number of frames in the acoustic context window, $w = 2c + 1$, is set to 19.

In Table 1, ACRFs results are reported for the Gaussian set trained using MLE criterion and the Gaussian set trained using MPE criterion. The ACRFs are estimated using CML criterion. The results show that ACRFs recognition performance is better than that HMMs recognition performance in terms of PER on TIMIT task. In addition, the ACRFs modeling based on activation functions trained using MPE criterion leads to improvements in some cases. For example, when the number of Gaussians in the acoustic model is 288 Gaussians (i.e. 2 mixtures per state), a 0.6 % absolute improvement is achieved over the results of the same Gaussians trained using MLE criterion. In the case of 2880 Gaussians based acoustic model, a 0.3 % absolute improvement is achieved over the results of the same Gaussians trained using MLE criterion. When the acoustic model complexity is high (i.e. 5760 Gaussians), the ACRFs based on discriminatively trained activation functions results did not outperform the results of ACRFs based on generatively trained activation functions.

6. Conclusions

In this paper, we present a method to construct ACRFs based on discriminatively trained activation functions. The estimated activation functions are used to construct a data driven, sparse, augmented space to increase discrimination. To estimate these activation functions, we train HMMs using MPE criterion. Hence, we ignore all mixture weights or transition matrices in an acoustic model and use the resultant Gaussians to construct the augmented space. The new derived Gaussians are no longer represent the dense regions in the observation space as in generative training. In some cases, the ACRFs modeling based on activation functions trained using MPE criterion leads to improvements over the ACRFs derived from generatively trained activation functions. When the number of Gaussians in the acoustic model is small, complexity is low, the improvements are bigger than high complexity models. Therefore, the results show that generative Gaussians work as well as the discriminative ones when the number of Gaussians are large. This is expected since ACRF is a discriminative method and using more discriminative features do not lead to big improvements for complex models.

7. References

[1] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001, pp. 282–289.

[2] Y. Hifny, "Conditional random fields for continuous speech recognition," Ph.D. dissertation, University Of Sheffield, 2006.

[3] Y. Hifny and S. Renals, "Speech recognition using augmented conditional random fields," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 2, pp. 354–365, 2009.

[4] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, "Hidden conditional random fields for phone classification," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1117–1120.

[5] M. Layton and M. Gales, "Augmented statistical models for speech recognition," in *Proc. IEEE ICASSP*, vol. 1, France, May 2006, pp. 129– 132.

[6] R. Prabhavalkar and E. Fosler-Lussier, "Backpropagation training for multilayer conditional random field based phone recognition," in *Proc. IEEE ICASSP*, vol. 1, France, March 2010, pp. 5534 – 5537.

[7] D. Yu and L. Deng, "Deep-structured hidden conditional random fields for phonetic recognition," in *Proc. INTERSPEECH*, 2010.

[8] G. Zweig, P. Nguyen, D. V. Compernolle, K. Demuynck, L. Atlas, P. Clark, G. Sell, M. Wang, F. Sha, H. Hermansky, D. Karakos, A. Jansen, S. Thomas, G. Sivaram, S. Bowman, and J. Kao, "Speech recognition with segmental conditional random ?elds: A summary of the jhu clsp summer workshop," in *Proc. IEEE ICASSP*, 2011.

[9] M. Gales, S. Watanabe, and E. Fosler-Lussier, "Structured discriminative models for speech recognition," *IEEE Signal Processing Magazine*, Nov 2012.

[10] Y. Hifny, S. Renals, and N. Lawrence, "A hybrid MaxEnt/HMM based ASR system," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 3017–3020.

[11] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, vol. EC-14, no. 3, pp. 326–334, 1965.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[13] K. Na, B. Jeon, D. Chang, S. Chae, and S. Ann, "Discriminative training of hidden Markov models using overall risk criterion and reduced gradient method," in *Proc. EUROSPEECH*, Madrid, Spain, September 1995, pp. 97–100.

[14] J. Kaiser, B. Horvat, and Z. Ka, "Overall risk criterion estimation of hidden Markov model parameters," *Speech Communication*, vol. 38, no. 3-4, pp. 383–398, 2002.

[15] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE ICASSP*, vol. 1, Orlando, FL, May 2002, pp. 105–108.

[16] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 2004.

[17] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P.Woodland, *The HTK Book, Version 3.4.1*. Cambridge University, March 2009. [Online]. Available: http://htk.eng.cam.ac.uk