SYSTEM COMBINATION WITH LOG-LINEAR MODELS

J. Yang, C. Zhang, A. Ragni, M. J. F. Gales and P. C. Woodland

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge, CB2 1PZ, UK
{jy308,cz277,ar527,mjfg,pcw}@eng.cam.ac.uk

ABSTRACT

Improved speech recognition performance can often be obtained by combining multiple systems together. Joint decoding, where scores from multiple systems are combined during decoding rather than combining hypotheses, is one efficient approach for system combination. In standard joint decoding the frame log-likelihoods from each system are used as the scores. These scores are then weighted and summed to yield the final score for a frame. The system combination weights for this process are usually empirically set. In this paper, a recently proposed scheme for learning these system weights is investigated for a standard noise-robust speech recognition task, AURORA 4. High performance tandem and hybrid systems for this task are described. By applying state-of-the-art training approaches and configurations for the bottleneck features of the tandem system, the difference in performance between the tandem and hybrid systems is significantly smaller than usually observed on this task. A log-linear model is then used to estimate system weights between these systems. Training the system weights yields additional gains over empirically set system weights when used for decoding. Furthermore, when used in a lattice rescoring fashion, further gains can be obtained.

Index Terms— Joint decoding, tandem system, hybrid system, log-linear model, structured SVM

1. INTRODUCTION

In general, different systems have various characteristics, make different errors, and are expected to provide complementary advantages. Thus, state-of-the-art speech recognisers typically utilise multiple systems to make ensemble decisions. Two system combination approaches called recogniser output voting error reduction (ROVER) [1] and confusion network combination (CNC) [2] are commonly used in speech recognition. The difference between these two is that ROVER uses the 1-best output, whereas CNC uses confusion networks [3]. In these schemes, multiple passes of decoding are required. In joint decoding [4], the systems to be combined can be trained separately possibly with different features (e.g. MFCC or filter bank) and training criteria (e.g. ML or MPE), but they share the same hidden Markov model (HMM) topology, and the frame level acoustic log-likelihoods from different systems are combined. Assume there are K different systems to be combined, given a speech frame (an observation) $\boldsymbol{o}_t$, in decoding the "log-likelihood" score corresponding to state s_i can be described as:

$$\mathcal{L}(\boldsymbol{o}_t|s_i) = \sum_{k=1}^{K} \eta_k \log p_k(\boldsymbol{o}_t|s_i) \tag{1}$$

where $\log p_k(\boldsymbol{o}_t|s_i)$ is the log-likelihood given by the kth system, and the scalar η_k is the corresponding combination weight. In work [4] combination of two forms of deep neural network (DNN) based systems were investigated, namely the tandem and hybrid systems [5, 6]. However, in [4] these combination weights are set empirically, and only system-dependent, which means the weights corresponding to different states for a system are set to be the same. Since in joint decoding the log-likelihoods from different systems are linearly combined with corresponding combination weights, it would be a natural extension by modelling these weights with a log-linear model and relaxing these weights to be state or phone dependent, which will be studied in this paper. In experiments, the phone dependent weights learnt by log-linear models[1] will be examined. In addition to joint decoding, decoding with log-linear models based on the segment level features derived from different systems [7] also will be studied. By using a lattice, decoding then can be operated in a lattice rescoring fashion [8].

2. LOG-LINEAR MODELS

In speech recognition the possible number of classes for an utterance can be exponentially large. For example, the possible number of classes for a 6-digit length utterance is 10^6. One solution to this problem is to segment the continuous speech into segments, and then classify each (independent) segment in an acoustic code breaking fashion [9]. Another solution is to introduce structure by breaking the sentence label down into sub-sentence units, such as words or phones with associated segmentation of the sentence. In a structured model such as the HMM, the structure of the class label is considered, and the parameters for any class (sentence) can be constructed from a common set of basic units [10]. In a structured discriminative model, the conditional distribution of the class label (sentence)

This work was supported in part by EPSRC Project EP/I006583/1 (Generative Kernels and Score Spaces for Classification of Speech) within the Global Uncertainties Programme, in part by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology), and in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. Supporting data for this paper is available at the https://www.repository.cam.ac.uk/handle/1810/253409 data repository.

[1] It is worth noting that only the segment level log-linear models will be used in experiments.

W given an input $\boldsymbol{O}$ can be can be described as:

$$P(W|\boldsymbol{O},\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta},\boldsymbol{O})} \sum_{\boldsymbol{\rho}\in\boldsymbol{P}_W} \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\Phi(\boldsymbol{O},W,\boldsymbol{\rho})\right) \quad (2)$$

where $\boldsymbol{\rho}$ is one possible segmentation of $\boldsymbol{O}$. $\mathcal{Z}(\boldsymbol{\eta},\boldsymbol{O})$ is the normalisation term that ensures $P(W|\boldsymbol{O},\boldsymbol{\eta})$ is a valid probability. The set $\boldsymbol{P}_W$ consists of all possible segmentations corresponding to the hypothesis W. Vector $\boldsymbol{\eta}$ is the parameter of the model and $\Phi(\boldsymbol{O},W,\boldsymbol{\rho})$ is the joint feature, which characterises the dependence between the input $\boldsymbol{O}$ and hypothesis W, and maps the input $\boldsymbol{O}$ with variable length to a fixed dimension [10]. The definition of the joint feature will be discussed in detail in the following section.

This type of structured discriminative model described in equation (2) is known as the conditional augmented (CAug) model [11] or segmental conditional random fields (SCRF) [12]. In this model, the summation over all possible segmentations makes training complicated. Alternatively, a variant of Viterbi training [13] could be used, where the most likely segmentation $\boldsymbol{\rho}$ from the HMM can be used instead of summing over all possible segmentations [14]. Then the structured discriminative model described in (2) can be approximated as a *log-linear model*:

$$P(W|\boldsymbol{O},\boldsymbol{\eta}) \approx \frac{1}{\mathcal{Z}(\boldsymbol{\eta},\boldsymbol{O})} \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\Phi(\boldsymbol{O},W,\boldsymbol{\rho})\right) \quad (3)$$

As an alternative to using the most likely segmentations from HMMs [15], optimal segmentations can be obtained from discriminative models [16].

3. FEATURE SPACE

Normally the features for structured discriminative models can be divided into two groups, namely the frame level and segment level features. In this work these two type of features based on log-likelihoods will be discussed.

Given one possible segmentation (or alignment) $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{|\rho|}$ which segments the sentence into sub-sentence units, the input utterance and sentence can be decomposed into $\boldsymbol{O} = \{\boldsymbol{O}_{(i)}\}_{i=1}^{|\rho|}$ and $W = \{w_i\}_{i=1}^{|\rho|}$, where $|\rho|$ is the number of segments. One general form of the joint feature $\Phi(\boldsymbol{O},W,\rho)$ based on K different systems can be described as [10, 14]:

$$\Phi(\boldsymbol{O},W,\rho) = \begin{bmatrix} \sum_{i=1}^{|\rho|} \boldsymbol{\phi}_1(\boldsymbol{O}_{(i)},\rho_i) \\ \vdots \\ \sum_{i=1}^{|\rho|} \boldsymbol{\phi}_K(\boldsymbol{O}_{(i)},\rho_i) \\ \boldsymbol{\phi}_{\text{lg}}(W,\rho) \end{bmatrix}; \; \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_K \\ \boldsymbol{\eta}_{\text{lg}} \end{bmatrix} \quad (4)$$

where $\boldsymbol{\phi}_k(\cdot)$ denotes the acoustic features for one segment from the kth system, and $\boldsymbol{\eta}_k$ are the corresponding weights. $\boldsymbol{\phi}_{\text{lg}}(\cdot)$ denotes the language features which provide pronunciation probabilities, word statistics, etc., and $\boldsymbol{\eta}_{\text{lg}}$ are the corresponding weights.

For the segment level features, various forms can be used, e.g. the log-likelihood [16] and derivative features [17]. In this work, the (segment level) acoustic features from the kth system are based on log-likelihoods:

$$\boldsymbol{\phi}_k(\boldsymbol{O}_{(i)},\rho_i) = \begin{bmatrix} \delta(w_i,v_1)\mathcal{L}_k(\boldsymbol{O}_{(i)}) \\ \vdots \\ \delta(w_i,v_L)\mathcal{L}_k(\boldsymbol{O}_{(i)}) \end{bmatrix} \quad (5)$$

where $\{v_l\}_{l=1}^{L}$ denotes all possible sub-sentence units (such as tri-phones) in the dictionary. $\mathcal{L}_k(\boldsymbol{O}_{(i)})$ is the log-likelihood given by the tri-phone model from the kth system:

$$\mathcal{L}_k(\boldsymbol{O}_{(i)}) = \log p_k(\boldsymbol{O}_{(i)}|w_{i-1}\text{-}w_i\text{+}w_{i+1}) \quad (6)$$

For robust parameter estimation it is important to tie the parameters $\boldsymbol{\eta}_k$. If these parameters are tied at the global central phone level then yields phone dependent weights. Alternatively, it is possible to use automatic approaches based on phonetic decision trees [18]. In the definition of the joint feature (4), the language feature $\boldsymbol{\phi}_{\text{lg}}(\cdot)$ for the segment level features can be defined as:

$$\boldsymbol{\phi}_{\text{lg}}(W,\rho) = \log P(W); \quad \boldsymbol{\eta}_{\text{lg}} = \eta_l \quad (7)$$

where $P(W)$ is given by the language model. The corresponding weight $\boldsymbol{\eta}_{\text{lg}}$ described in (4) is a scalar η_l, which functions as scaling language model. In this work, the log-linear model based on the segment level features is called the *segment level log-linear model*. More analysis of the segment level features can be found in our previous work [7]. In experiments, only the segment level features will be examined. Since the log-linear model based on the frame level features is related to joint decoding[2], a brief discussion of the frame level features is also given in the rest of this section.

For the frame level features, normally another level of hidden information is introduced [10, 19], namely the state sequence $\boldsymbol{\theta} = \{\theta_t\}_{t=1}^{T}$ corresponding to the input utterance $\boldsymbol{O} = \{\boldsymbol{o}_t\}_{t=1}^{T}$, where $\boldsymbol{o}_t$ is the tth frame of the utterance. Then for one segment the (frame level) acoustic features from the kth system can be described as:

$$\boldsymbol{\phi}_k(\boldsymbol{O}_{(i)},\rho_i) = \begin{bmatrix} \sum_{t\in\{\rho_i\}} \delta(\theta_t,s_1)\mathcal{L}_k(\boldsymbol{o}_t) \\ \vdots \\ \sum_{t\in\{\rho_i\}} \delta(\theta_t,s_I)\mathcal{L}_k(\boldsymbol{o}_t) \end{bmatrix} \quad (8)$$

where $\{\rho_i\}$ denotes the indexes of the frames associated with the ith segment. $\{s_i\}_{i=1}^{I}$ denotes all possible states, and θ_t takes value from this set. $\mathcal{L}_k(\boldsymbol{o}_t)$ is the frame feature from the kth system. This feature can have various forms, e.g. the Gaussian sufficient statistics [19] and HMM mean and variance statistics [20]. One of the simplest forms for $\mathcal{L}_k(\boldsymbol{o}_t)$ is a log-likelihood:

$$\mathcal{L}_k(\boldsymbol{o}_t) = \log p_k(\boldsymbol{o}_t|\theta_t) \quad (9)$$

For the frame level features, in the definition of the joint feature (4), the language features $\boldsymbol{\phi}_{\text{lg}}(\cdot)$ can be described as:

$$\boldsymbol{\phi}_{\text{lg}}(W,\rho) = \begin{bmatrix} \log P(\boldsymbol{\theta}) \\ \log P(W) \end{bmatrix}; \quad \boldsymbol{\eta}_{\text{lg}} = \begin{bmatrix} \eta_s \\ \eta_l \end{bmatrix} \quad (10)$$

where $\log P(\boldsymbol{\theta})$ is given by the state transition probabilities. As discussed in section 1, in joint decoding different systems share the same HMM topology, here the state transition probabilities are shared with systems. Analogous to the weight η_l that scales language models, η_s is used to scale the transition probabilities. In order to ensure valid state transition probabilities, this weight can be fixed to one. In this work, the log-linear model based on the frame level features is called the *frame level log-linear model*.

[2]The relationship will be discussed in the decoding section.

4. TRAINING CRITERIA

In the previous sections, the log-linear model and the forms of features were discussed. For a log-linear model, it can be trained with various training criteria, and the most commonly used ones will be discussed in this section. Given a training set consisting of utterance and reference pairs $\mathcal{D} = \{(\boldsymbol{O}_n, W_n)\}_{n=1}^N$, the log-linear model parameters can be estimated by maximising the conditional maximum likelihood (CML) training criterion. Normally, prior information of the model parameters is available, e.g. the weights set empirically in joint decoding as described in (1). Thus a prior can be introduced in training. A Gaussian prior, $\log p(\boldsymbol{\eta}) = \log \mathcal{N}(\boldsymbol{\mu}_\eta, C\boldsymbol{I}) \propto -\frac{1}{C}||\boldsymbol{\eta} - \boldsymbol{\mu}_\eta||^2 + \text{Constant}$, is usually used [14]. Then the CML criterion can be expressed as:

$$\mathcal{F}_{\text{CML}}(\boldsymbol{\eta}) = -\frac{1}{C}||\boldsymbol{\eta} - \boldsymbol{\mu}_\eta||^2 + \sum_{n=1}^{N} \log P(W_n|\boldsymbol{O}_n, \boldsymbol{\eta}) \tag{11}$$

Large margin training has been studied in speech recognition, and the state-of-the-art performance can be achieved with this criterion [7, 10], given its advantage in generalisation [21]. Analogous to CML estimation, a Gaussian prior is also introduced in training. The parameters of the log-linear model can be estimated by minimising:

$$\mathcal{F}_{\text{LM}}(\boldsymbol{\eta}) = \frac{1}{C}||\boldsymbol{\eta} - \boldsymbol{\mu}_\eta||^2 + \sum_{n=1}^{N} \Bigg[\max_{W \neq W_n} \Big\{ \mathcal{L}(W, W_n) - \log\Big(\frac{P(W_n|\boldsymbol{O}_n, \boldsymbol{\eta})}{P(W|\boldsymbol{O}_n, \boldsymbol{\eta})}\Big) \Big\} \Bigg]_+ \tag{12}$$

where $\mathcal{L}(W, W_n)$ is the loss function, which measures how different the hypothesis W and the reference W_n are, e.g. the loss function can be computed between phone sequences. The number of all possible hypotheses for an utterance is exponentially large, but a lattice can be used to limit the search space of hypotheses. Thus, in practice the best competing hypothesis W for each instance can found in a denominator lattice [22]. Substituting the definition of the log-linear model (3) into criterion (12), the denominator term of the log-linear model can be cancelled out. Then large margin criterion (12) can be further written as minimising:

$$\mathcal{F}_{\text{LM}}(\boldsymbol{\eta}) = \frac{1}{C}||\boldsymbol{\eta} - \boldsymbol{\mu}_\eta||^2 + \sum_{n=1}^{N} \Bigg[\max_{W,\rho \neq W_n,\rho_n} \Big\{ \mathcal{L}(W, W_n) + \boldsymbol{\eta}^{\mathsf{T}}\Phi(\boldsymbol{O}_n, W, \rho) \Big\} - \boldsymbol{\eta}^{\mathsf{T}}\Phi(\boldsymbol{O}_n, W_n, \rho_n) \Bigg]_+ \tag{13}$$

As discussed in section 2, the most likely segmentation is considered rather than summing over all possible segmentations. Similarly, when finding the best competing hypothesis W, only one corresponding segmentation ρ is considered. As described in equation (13), the best competing hypothesis and segmentation pair (W, ρ) is found over all possible labels and segmentations except the reference with the corresponding segmentation (W_n, ρ_n), where ρ_n is the most likely segmentation obtained from the 1-best output of joint decoding as discussed in section 2. Equation (13) is the training criterion of the structured SVM [8], and it can be efficiently solved by using the cutting-plane algorithm [23].

5. DECODING

Different training criteria were discussed in the previous section. By using these training criteria, the optimal model parameters $\hat{\boldsymbol{\eta}}$ can be estimated. Given the optimal parameters $\hat{\boldsymbol{\eta}}$ and an input $\boldsymbol{O}$, decoding with the log linear model defined in (3) can be described as:

$$\hat{W} = \arg\max_{W} P(W|\boldsymbol{O}, \hat{\boldsymbol{\eta}}) \tag{14}$$

As discussed in section 2, in the log-linear model the most likely segmentation is used. Then decoding yields both the optimal word sequence $\hat{W}$ and segmentation $\hat{\rho}$ [16]:

$$\hat{W} = \arg\max_{W} \Big\{ \max_{\boldsymbol{\rho}} \hat{\boldsymbol{\eta}}^{\mathsf{T}}\Phi\big(\boldsymbol{O}, W, \rho\big) \Big\} \tag{15}$$

This is equivalent to decoding with structured SVM [10]. Normally, the hypothesis and corresponding segmentation can be found from a lattice, which gives the information about possible word sequences and segmentations in a reasonable size. Then, decoding with the log-linear model based on the segment level or frame level features becomes a lattice rescoring approach. In experiments, only decoding with the log-linear model based on the segment level features will be examined.

When the segment level features are used, the log-linear model parameters $\hat{\boldsymbol{\eta}}$ could be considered as phone dependent acoustic model scales [24]. Then, these phone dependent weights can be applied to joint decoding, and this approach will be examined in experiments. In the rest of this section, the relationship between decoding with frame level log-linear models and joint decoding will be discussed.

When the frame level features are used, substituting the definitions of the frame level features (4), (8) and (10) in, decoding described in (15) becomes[3]:

$$\hat{W} = \arg\max_{W} \Big\{ \max_{\boldsymbol{\theta}} \Big(\eta_s \log P(\boldsymbol{\theta}) + \eta_l \log P(W) + \sum_{t=1}^{T}\sum_{i=1}^{I} \delta(\theta_t, s_i) \sum_{k=1}^{K} \eta_{k,i} \log p_k(\boldsymbol{o}_t|\theta_t) \Big) \Big\} \tag{16}$$

where $P(\boldsymbol{\theta})$ is the state transition probability and $P(W)$ is the probability given by language model. t is the index for frames, i is the index for states, and k is the index for systems. $\eta_{k,i}$ is the ith elements of the parameters $\boldsymbol{\eta}_k$ corresponding to the kth system described in (4). For the tth frame, in decoding as described in (16), the score computed for state s_i can be described as:

$$\mathcal{L}(\boldsymbol{o}_t|s_i) = \sum_{k=1}^{K} \eta_{k,i} \log p_k(\boldsymbol{o}_t|s_i) \tag{17}$$

This is the same as the combined score used in joint decoding described in (1), but with more general state dependent weights $\boldsymbol{\eta}_k^{\mathsf{T}} = [\eta_{k,1}, \ldots, \eta_{k,I}]$ for the kth system. When $\eta_s = 1$, decoding with log-linear models becomes HMM Viterbi decoding, and it is equivalent to standard joint decoding described in section 1.

6. EXPERIMENTS

The experiments were conducted on AURORA 4 corpus, which is a noise-corrupted medium vocabulary corpus. In experiments the multi-condition training set was used. This set is artificially corrupted from the clean training set with different noise and channel

[3] Given the state sequence $\boldsymbol{\theta}$, for each frame the associated state of the tri-phone model is known, namely the phone boundary is known. Thus segmentation information is embedded in the state sequence. Then, given the optimal state sequence $\boldsymbol{\theta}$, the optimal segmentation $\boldsymbol{\rho}$ is known.

conditions. The test set is based on the development set of 1992 November NIST evaluation, and it is artificially corrupted by using 6 types of noise under 2 channel conditions. The test set consists of 4 sets: A, B, C and D. Set A is clean, set B has 6 types of additive noise, set C has channel distortion, and set D has both additive noise and channel distortion.

6.1. Tandem, Hybrid, and Joint Decoding Baseline Systems

A tandem and hybrid system combination was studied in this work ($K = 2$). Both systems were trained and decoded using HTK V3.5 [13, 25]. $L2$ regularisation was used for cross entropy (CE) training with a scaling factor of 0.001. Parameter updates were averaged over a mini-batch with 200 frames and were smoothed by adding a "momentum" term of 0.9 times the previous updates. Single epoch discriminative pre-training was performed with a learning rate of 1.0×10^{-3} [6]. The "fine-tuning" stage used a modified NewBob learning rate scheduler [25], with the initial learning rate 2.0×10^{-3} and a minimum epoch number 16.

The tandem and hybrid systems were built with 13d PLP and 40d log-Mel filter bank (FBK) coefficients. A triphone state set with 3063 tied-states produced by the decision tree tying approach was used by each system [25]. A MPE trained GMM-HMM system with 52d PLP+Δ+Δ^2+Δ^3 features were used to produce the frame-to-state alignments for CE based bottleneck (BN) DNN training. The BN DNN structure is $720 \times 2000^4 \times 39 \times 2000 \times 3064$, whose input vector is formed by stacking 80d FBK+Δ features according to a *context shift set* $\mathbf{c} = [-4, +4]$ (to stack the current frame with 4 frames in its left and right contexts) [25]. The 39d BN feature vectors were de-correlated with semi-tied covariance (STC) [26] matrix and then combined with 52d to 39d heteroscedastic linear discriminant analysis (HLDA) projected PLP+Δ+Δ^2+Δ^3 [27].

The MPE trained speaker independent tandem system has 32 Gaussian components for each of the 3 silence states and 16 Gaussian components per state for the others. For the hybrid system, the DNN acoustic model with sigmoid activation function, whose structure is $720 \times 2000^5 \times 3066$, was initially trained with CE using the alignments generated by the tandem system. Once the hybrid system was trained, it was refined with MPE based sequence discriminative training [25], with a fixed learning rate of 1.0×10^{-5} for 6 epochs.

To form the joint decoding system, the MPE traned tandem and hybrid systems were combined using HTK joint decoder [25] based on frame-level log-likelihood linear combination (as described in equation (1)) and system dependent weights $\{0.2, 1.0\}$.

As tabulated in Table 1, the baseline results of the MPE trained tandem and hybrid systems are considerably better than the previously published numbers [7, 28, 29]. The performance gains are mainly due to the use of 40d FBK rather than 24d FBK features, regularised discriminative pre-training rather than generative pre-training, as well as the MPE discriminative sequence training (based on a high performance CE hybrid system with 11.64% WER on average). In joint decoding, 2% relative WER performance gain was achieved over the hybrid system, from 11.24% to 11.04%.

6.2. Log-linear Model Combination

In experiments, only segment level log-linear models (LLM) described in sections 2 and 3 were examined. Two training criteria were used in training the log-linear models, namely the conditional maximum likelihood (CML) and large margin (LM) training criteria. In training, as described in section 4, a Gaussian prior over the log-linear model parameters $\boldsymbol{\eta}$ was used. The mean of the prior is set to be the combination weights for the joint decoding baseline system, namely the weights corresponding to the hybrid and tandem systems are 1.0 and 0.2 respectively. With this prior, the log-linear model can start with a good configuration by using a small variance (small value of C), and the optimal configuration can be obtained by gradually increasing C.

System	Criterion	Test Set WER(%)				Avg.
		A	B	C	D	
Tandem	MPE	4.78	7.63	8.93	19.14	12.45
Hybrid		3.75	6.70	7.68	17.62	11.24
CNC	–	3.87	6.76	7.45	17.17	11.06
Joint	Empirical	3.79	6.47	7.86	17.34	11.04
	CML	3.74	6.47	7.73	17.19	10.96
	LM	3.59	6.50	7.21	17.05	10.87
LLM	Empirical	3.74	6.57	7.88	17.12	10.98
	CML	3.66	6.56	7.88	17.06	10.94
	LM	3.64	6.56	7.04	16.83	10.79

Table 1. *The WER performance on AURORA 4 corpus*

In Table 1, "CNC" denotes the confusion network combination [2]. "Joint" means joint decoding as discussed in section 1. "LLM" indicates decoding with the segment level log-linear model in a lattice[4] rescoring fashion as described in section 5. "CML" and "LM" means the (phone dependent) weights were estimated by the segment level log-linear models with CML and LM training criteria.

As tabulated in Table 1, joint decoding (where only one pass of decoding is required) outperforms CNC. As described in the 3rd block of Table 1, compared with joint decoding using the empirically set weights, when the weights were estimated by the segment level log-linear model (with the CML or LM criterion), performance gains can be achieved. As given in the 4th block of Table 1, when applying these (CML or LM trained) weights to log-linear model decoding that operates in a lattice rescoring fashion, further performance gains can be achieved. The possible reason is that in joint decoding different models are constrained to be state synchronised. By introducing lattices, the constraint is relaxed to phone level and the log-likelihoods for each phone are allowed to be computed with forward algorithm. Another reason is that in joint decoding, these phone dependent weights are treated as acoustic model scales, whereas, by using lattices which provide segmentation information, decoding with the segment level log-linear model can result in the form of lattice rescoring [14]. As tabulated in the bottom line of Table 1, when the weights were estimated with LM criterion, the best average WER 10.79% can be achieved by the segment level log-linear model, with 2% relative performance gain over joint decoding using the empirically set weights (the 4th line).

7. CONCLUSION

In this paper, the state-of-the-art training approaches and configurations for the tandem, hybrid and joint decoding systems have been detailed on AURORA 4 task. By using the (phone dependent) weights learnt by the segment level log-linear models, decoding with log-linear models in a lattice rescoring fashion yields good performance gains over joint decoding with empirically set weights. These phone dependent acoustic model weights also have been successfully applied to joint decoding. This motivates training of the state dependent acoustic model weights for joint decoding as the future work.

[4] In all experiments, lattices are generated by the joint decoding system with empirically set weighs.

8. REFERENCES

[1] Jonathan G Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 1997, pp. 347–354.

[2] Gunnar Evermann and PC Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proceedings of Speech Transcription Workshop*. Baltimore, 2000, vol. 27.

[3] Mark Gales and Steve Young, "The application of hidden Markov models in speech recognition," in *Foundations and Trends in Signal Processing*, 2007, pp. 195–304.

[4] H. Wang, A. Ragni, M. Gales, K. Knill, P. Woodland, and C. Zhang, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *Proceedings of Interspeech*, 2015.

[5] Hynek Hermansky, Daniel W Ellis, and Shantanu Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proceedings of ICASSP*. IEEE, 2000, vol. 3, pp. 1635–1638.

[6] G. Hinton, L. Deng, D. Yu, D. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, pp. 2–17, Nov. 2012.

[7] R. C. van Dalen, J. Yang, H. Wang, A. Ragni, C. Zhang, and M. J. F. Gales, "Structured discriminative models using deep neural-network features," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015.

[8] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 16, pp. 1453–1484, 2005.

[9] Veera Venkataramani, Shantanu Chakrabartty, and William Byrne, "Support vector machines for segmental minimum bayes risk decoding of continuous speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003, pp. 13–18.

[10] Shi-Xiong Zhang, *Structured Support Vector Machines for Speech Recognition*, Ph.D. thesis, University of Cambridge, 2014.

[11] Martin Layton, *Augmented statistical models for classifying sequence data*, Ph.D. thesis, University of Cambridge, 2006.

[12] Geoffrey Zweig and Patrick Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2009, pp. 152–157.

[13] S. Young, G. Evermann, M. Gales, T. Hain., D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. Woodland, and C. Zhang, *The HTK book (for HTK version 3.5)*, Cambridge University Engineering Department, Cambridge, UK, 2015.

[14] Anton Ragni, *Discriminative models for Speech Recognition*, Ph.D. thesis, University of Cambridge, 2013.

[15] Mark Gales, "Discriminative models for speech recognition," in *Information Theory and Applications Workshop, 2007*. IEEE, 2007, pp. 170–176.

[16] Shi-Xiong Zhang and Mark Gales, "Structured SVMs for automatic speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, pp. 544–555, 2013.

[17] Anton Ragni and MJF Gales, "Derivative kernels for noise robust asr," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2011, pp. 119–124.

[18] Anton Ragni and Mark John Francis Gales, "Structured discriminative models for noise robust continuous speech recognition," in *Proceedings of ICASSP*. IEEE, 2011, pp. 4788–4791.

[19] Asela Gunawardana, Milind Mahajan, Alex Acero, and John C Platt, "Hidden conditional random fields for phone classification.," in *Proceedings of Interspeech*, 2005, pp. 1117–1120.

[20] Georg Heigold, Ralf Schlüter, and Hermann Ney, "On the equivalence of Gaussian HMM and Gaussian HMM-like hidden conditional random fields.," in *Proceedings of Interspeech*, 2007, pp. 1721–1724.

[21] Vladimir N. Vapnik, *Statistical learning theory*, vol. 1, Wiley New York, 1998.

[22] Daniel Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, University of Cambridge, 2003.

[23] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu, "Cutting-plane training of structural SVMs," *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.

[24] Björn Hoffmeister, Ruoying Liang, Ralf Schlüter, and Hermann Ney, "Log-linear model combination with word-dependent scaling factors.," in *Proceedings of Interspeech*, 2009, pp. 248–251.

[25] C. Zhang and P.C. Woodland, "A general artificial neural network extension for HTK," in *Proceedings of Interspeech*, Dresden, Germany, 2015.

[26] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[27] X. Liu, M.J.F. Gales, and P.C. Woodland, "Automatic complexity control for HLDA systems," in *Proceedings of ICASSP*, Hong Kong, 2003.

[28] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings of ICASSP*. IEEE, 2013, pp. 7398–7402.

[29] Chao Weng, Dong Yu, Shigetaka Watanabe, and Biing-Hwang Fred Juang, "Recurrent deep neural networks for robust speech recognition," in *Proceedings of ICASSP*. IEEE, 2014, pp. 5532–5536.