SEGMENT-LEVEL TRAINING OF ANNS BASED ON ACOUSTIC CONFIDENCE MEASURES FOR HYBRID HMM/ANN SPEECH RECOGNITION

S. Pavankumar Dubagunta[1,2] *and Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

We show that confidence measures estimated from local posterior probabilities can serve as objective functions for training ANNs in hybrid HMM based speech recognition systems. This leads to a segment-level training paradigm that overcomes the limitation of frame-level updates ignoring the sequence structure in speech. We propose measures that train at the state and phone segment levels, while still decoding in the conventional framework. Experimental results on multiple corpora show that such trainings not only yield better systems in terms of performance, but also give additional improvements with sequence discriminative training. These techniques generalise across front-ends and model architectures, and efficiently handle the effect of segment duration variations on the ANN training.

Index Terms— Speech recognition, confidence measures, local posterior probability, segment-level training.

1. INTRODUCTION

An automatic speech recognition (ASR) system converts speech signals into sequences of words or text. In hidden Markov model (HMM) based ASR, the *likelihood* of an HMM state q_t at the time frame t, labelled l^i, is estimated [1] as:

$$\begin{aligned} p(\mathbf{x}_t|q_t = l^i) &= \sum_{d=1}^{D} p(\mathbf{x}_t, a^d|q_t = l^i) \\ &= \sum_{d=1}^{D} P(a^d|q_t = l^i) \cdot p(\mathbf{x}_t|a^d, q_t = l^i) \quad (1) \\ &= \sum_{d=1}^{D} P(a^d|q_t = l^i) \cdot p(\mathbf{x}_t|a^d) \,, \quad (2) \end{aligned}$$

where $\mathbf{x}_t$ denotes the acoustic feature observation at t, $l^i \in \{1, \ldots, I\}$ and $\{a^d\}_{d=1}^{D}$ denotes a set acoustic units. Eqn. (2) results from the assumption that $\mathbf{x}_t \perp\!\!\!\perp q_t|a^d$. In the case of a context dependent subword unit based ASR system, I is the number of context-dependent subword units; D is the number of clustered context-dependent states; and the vector $[P(a^d|q_t = l^i)]_{d=1}^{D}$ is either a Kronecker delta distribution or a soft distribution depending upon whether the relationship between a^d and state $q_t = l^i$ is deterministic or probabilistic [1]. In standard HMM-based ASR systems this relationship is deterministic given the state tying decision tree, i.e. if $l^i \mapsto a^{d'}$ then $P(a^{d'}|q_t = l^i) = 1$ and $P(a^d|q_t = l^i) = 0 \; \forall d \neq d'$. $p(\mathbf{x}_t|a^d)$ can be estimated either using Gaussian mixture models (GMM) or using artificial neural networks (ANN). In the case of ANNs, $p(\mathbf{x}_t|a^d)$ is estimated as a scaled-likelihood $p_{sl}(\mathbf{x}_t|a^d)$ [2]:

$$p_{sl}(\mathbf{x}_t|a^d) = \frac{p(\mathbf{x}_t|a^d)}{p(\mathbf{x}_t)} = \frac{P(a^d|\mathbf{x}_t)}{P(a^d)}, \quad (3)$$

where $P(a^d|\mathbf{x}_t)$ denotes the posterior probability of the acoustic unit a^d estimated by the ANN and $P(a^d)$ is its prior probability.

The focus of this paper lies in the training of the ANNs to estimate $P(a^d|\mathbf{x}_t)$. The ANN can be trained using embedded Viterbi expectation-maximisation (EM) algorithm. In the expectation step (E-step), given the current neural network, an alignment between the HMM state sequences and the acoustic feature sequences is obtained. In the maximisation step (M-step), given the alignment, a new neural network is trained. In practice, to reduce the training time, the alignments are typically obtained using an HMM/GMM system and the M-step is carried out once [3, 4].

Although the alignment is obtained by imposing a sequence structure, the ANN is trained using an individual frame-level discriminative criterion, viz. cross-entropy (CE). This training criterion corresponds to a maximum mutual information (MMI) estimation of parameters in terms of classifying phones [5]. However, this may be sub-optimal since the sequence structure in the data is being ignored. One class of methods which can address this limitation is segmental models [6], where the HMM states emit segments instead of frames. These ideas have been used in ANN- and deep learning based models. Such methods often depend on the availability of segment boundaries in the data, and thus require an additional complexity to determine and handle variable length segments both during training and decoding. We mention a few examples among numerous works in the literature here. Austin et al. converted segments into fixed length segments by sampling the segments linearly [7]. This requires an additional rescoring process during decoding after the first pass, since an initial segmentation is unavailable during real-time testing. Abdel-Hamid et al. use similar sampling methods to carry out training, but expensively loop over multiple possible segment boundaries during decoding [8]. Zweig and Nguyen used a conditional random field based backend to combine outputs at multiple segment levels [9]. Kong et al. used a recurrent architecture [10] and Beck et al. used an encoder-decoder based framework [11]. Another class of methods that handle segments of speech together are sequence discriminative training (SDT) [12, 13] methods, where the training objectives are computed at sequence levels, while keeping the model complexity unaltered.

In this paper we first establish a link between the estimation of linguistic unit level confidences using $P(a^d|\mathbf{x}_t)$ [14, 15] and

This work was funded by the Hasler Foundation through the project Flexible Linguistically-guided Objective Speech aSsessment (FLOSS). The authors gratefully thank them for their financial support and for a fruitful collaboration. The authors thank Alexandre Nanchen for the help with Mediaparl language models, and Suraj Srinivas for the productive discussions.
e-mail: (see http://www.idiap.ch/en/people/directory).

the training of neural networks. Through this link we propose a segment-level training paradigm that requires no architectural changes or sophistication, and can be envisaged as a maximisation of segment- or linguistic unit level confidences. In other words, it can be viewed as the maximisation of the match between linguistic units and segments of acoustic feature observations. Through experimental studies on multiple corpora, we show that the proposed training methods lead to significantly better systems than using frame level cross entropy criterion. Furthermore, we also show that these gains are sustained or boosted further with sequence discriminative training.

The remainder of the paper is organised as follows. Section 2 presents the proposed method of segment level training. Section 3 describes the experimental setup and gives the results. Section 4 presents an analysis of the proposed approach. Section 5 finally concludes the paper, indicating the future directions.

2. PROPOSED SEGMENTAL TRAINING APPROACH

In ASR related applications, confidence measures are used to measure how well an acoustic observation sequence $\mathbf{X} = (\mathbf{x}_t)_{t=1}^{T} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$ and a word hypothesis $W = (w_r)_{r=1}^{R}$ match, given the trained parameters of the system. In a similar vein, the training of the ANN for ASR can be posed as finding the parameters that maximise the match between between $\mathbf{X}$ and W. In both the cases, matching $\mathbf{X}$ and W is a common factor; where higher the confidence, better the match. Given this understanding, in this section we show that confidence measures based on "local posterior" probability estimates $P(a^d|\mathbf{x}_t)$ can naturally serve as objective functions for a segment-level training of the ANNs.

2.1. Segment-level confidence estimation from local posteriors

Let $W = (w_r)_{r=1}^{R}$ constitute a sequence of phones $(ph_k)_{k=1}^{K}$, and further constitute a sequence of sub-phonemic HMM states $(s_j)_{j=1}^{J}$ as defined by the topology. In the framework of *acceptor HMMs* [16, 2], various confidence measures based on local posterior probability estimates have been proposed. Specifically, given an alignment between $\mathbf{X}$ and W and the local posterior probability estimates, one of the methods to estimate the HMM state level confidence $CM(s_j)$ is by rescoring the state segment s_j as

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} \log\left(P(q_t = l^j|\mathbf{x}_t)\right)}{e(s_j) - b(s_j) + 1}, \tag{4}$$

where l^j is its label, and $b(s_j)$ and $e(s_j)$ denote its beginning and end frames respectively. This is computed, given the one-to-one map between the state l^j and the set of acoustic units $\{a^d\}_{d=1}^{D}$. In other words, if $l^j \mapsto a^{d'}$ then

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} \log\left(P(a^{d'}|\mathbf{x}_t)\right)}{e(s_j) - b(s_j) + 1}. \tag{5}$$

A word level confidence $wCM(w_r)$ for the word w_r constituting the state sequence $(s_{j+m})_{m=1}^{M_{w_r}}$ can be further estimated as [14]

$$wCM(w_r) = \frac{1}{M_{w_r}} \sum_{m=1}^{M_{w_r}} CM(s_{j+m}), \tag{6}$$

where M_{w_r} is the number of states in w_r.

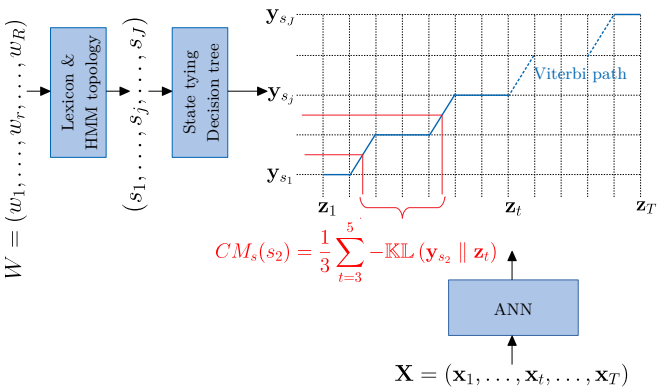

Fig. 1. Estimating state confidences from local posterior probabilities.

Let $y_{s_j}^d = P(a^d|q_t = l^j)$; then the vector $\mathbf{y}_{s_j} = (y_{s_j}^d)_{d=1}^{D}$ describes the mapping from s_j to $\{a^d\}_{d=1}^{D}$. As illustrated in Fig. 1, this mapping is typically defined by the state tying decision tree. In other words, the sequence $(s_j)_{j=1}^{J}$ that corresponds to a word hypothesis is mapped to $\mathbf{Y} = \left(\mathbf{y}_{s_j}\right)_{j=1}^{J}$. Similarly, let $z_t^d = P(a^d|\mathbf{x}_t)$; then the vector $\mathbf{z}_t = (z_t^d)_{d=1}^{D}$ denotes the output of the ANN at the time frame t and we can define the sequence $\mathbf{Z} = (\mathbf{z}_t)_{t=1}^{T}$ that corresponds to an acoustic observation. Without loss of generality, the estimation of confidence by rescoring can be expressed as a matching of the two posterior probability sequences $\mathbf{Y}$ and $\mathbf{Z}$ with a local cost based on Kullback-Leibler divergence $\mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right)$. More precisely,

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} -\mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right)}{e(s_j) - b(s_j) + 1}. \tag{7}$$

It can be verified that, as $\mathbf{y}_{s_j}$ is a Kronecker delta distribution given $s_j \mapsto d'$, $\mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right)$ reduces to *cross entropy* $-\log\left(P(a^{d'}|\mathbf{x}_t)\right)$.

It is worth mentioning that Eqn. (7) can be generalised further to the case when $\mathbf{y}_{s_j}$ is a soft distribution, as computing the KL-divergence between two probability distributions is equivalent to hypothesis testing [17, 18]. Indeed such confidence measures have been employed earlier for utterance verification [19] and for non-native speech assessment [20] tasks.

2.2. Segment-level training of the ANNs based on confidence measures

Given the segmentation of the training data, the ANN training is treated as a separate classifier training, by one hot encoding of the targets and minimising a frame level cross entropy criterion

$$\begin{aligned} E_f(t) &= \mathbb{KL}\left(\delta_d \parallel \mathbf{z}_t\right) \\ &= -\log\left(P(a^d|\mathbf{x}_t)\right), \end{aligned} \tag{8}$$

where δ_d is a Kronecker delta distribution based on a one-hot encoding and $\mathbf{z}_t$ is the output of the ANN. From this perspective, given the pairs of input features and their target classes as tuples, there is no difference in the training mechanism whether one wants to classify

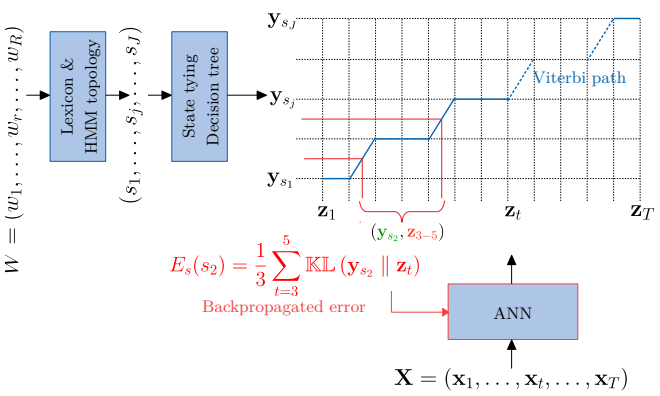

Fig. 2. Training from state level confidence scores.

phones, speakers, images, text or so on. This is a non-segmental way to train ANNs.

On the contrary, given the understanding from Section 2.1, the ANN training for hybrid HMM/ANN ASR can be formulated as finding the parameters that increase the match between the observation sequences and the sequence of states or segments. More precisely, as illustrated in Fig. 2, the error function can be based on rescoring of the segments, i.e. based on confidence measures. It is important to mention that whilst the notion of one-hot-encoding of the targets comes from a pattern classification point of view, in our formulation one-hot-encoding results from the one-to-one mapping between the states and $\{a^d\}_{d=1}^{D}$. As discussed earlier the targets can be soft, i.e. the map between the states and $\{a^d\}_{d=1}^{D}$ can be probabilistic. Furthermore, as shown earlier as well as in the literature, the cross entropy error criterion emerges from KL-divergence with the target distributions being Kronecker delta distributions [21]. In the case of soft targets, it corresponds to an additional entropy term of the target distributions, that remains constant with respect to the ANN parameters, and thus makes no difference in the training.

In the case where the segments represent HMM states, a state-level error function $E_s(s_j)$ that can be defined to minimise in a stochastic gradient descent training is

$$E_s(s_j) = -CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} \mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right)}{e(s_j) - b(s_j) + 1}, \qquad (9)$$

while in the case where the segments represent phone units, a phone-level error function $E_{ph}(ph_k)$ that is minimised can be based on Eqn. (6):

$$E_{ph}(ph_k) = \frac{1}{N_{ph_k}} \sum_{n=1}^{N_{ph_k}} E(s_{j+n}), \qquad (10)$$

where the phone ph_k constitues N_{ph_k} states: $(s_{j+1}, \ldots, s_{j+N_{ph_k}})$.

The decoding process remains unaltered, except that the priors $P(a^d)$ in Eqn. (3) are estimated from the state segment counts rather than from the frame label counts.

3. EXPERIMENTAL STUDY

In this section, we investigate the effect of the proposed segment level ANN training on ASR and phone recognition performances.

3.1. Data setup

We conducted ASR experiments on Mediaparl [22] and AMI [23] data sets. Mediaparl is a bilingual corpus containing data (debates) in both Swiss German (denoted as M-DE) and Swiss French (denoted as M-FR) which were recorded at the Valais parliament in Switzerland. Valais is a bilingual state which has both French and German speakers with a high variability in their local accents. We performed studies on both the M-DE and M-FR parts of the data set. We followed the protocols set in [22] for their data preparation, pronunciation lexicon selection and language model (LM) building. AMI is a meeting room corpus with data collected through an individual headset microphone (IHM), as well as a microphone array. We conducted the studies on the IHM data set. We conducted phone recognition studies on TIMIT corpus [24]. We followed the standard Kaldi protocols for AMI and TIMIT. Table 1 provides a description of the experimental setup for all the data sets.

Table 1. Experimental setup on various corpora.

	AMI	M-DE	M-FR	TIMIT
Training hours	77.3	14.5	16.1	3.1
Phone set count	176	57	38	48
Vocabulary size	52.5k	16.7k	12.4k	48
LM order	3-gram	2-gram	2-gram	2-gram

3.2. Systems

We built ASR systems using Kaldi toolkit and Keras/Tensorflow tools. We used 39 dimensional Mel frequency cepstral coefficients (MFCC), $C_0 - C_{12} + \Delta + \Delta\Delta$, as the acoustic feature observations. AMI and TIMIT used the default speaker-level cepstral mean and variance normalisation (CMVN) in Kaldi setup, while M-DE and M-FR used an utterance-level CMVN.

The alignments for the training of ANNs were obtained using Kaldi pipeline, by first building *mono-tri3* HMM/GMMs and then building subspace GMM (SGMM) systems, which operate in three passes for decoding and alignment. The number of clustered context-dependent states for AMI, M-DE, M-FR and TIMIT were 4490, 2282, 2265 and 2112 respectively. The alignments for each data set was obtained from its corresponding SGMM system. For AMI, it is worth mentioning that the SGMM system development and the subsequent ANN training were carried out on the 70.2 hour subset of data with *clean* segmentation.

For each data set, we trained three deep neural networks (DNNs) corresponding to the three error functions E_f, E_s and E_{ph}. All the DNNs had three hidden layers with 1024 units with rectified linear activations in each hidden layer. The input to the DNNs were 13 dimensional MFCCs with five frames each in the preceding and the following context and with $\Delta + \Delta\Delta$, i.e. 429 dimensional feature input. The training was based on stochastic gradient descent with a decaying learning rate. Post this training, we also used a standard sequence discriminative training, viz. state-level minimum Bayes risk (sMBR), for AMI, M-DE and M-FR corpora.

3.3. Results

Table 2 shows the word error rates (WER) on AMI, M-DE and M-FR corpora and phoneme error rate (PER) for TIMIT corpus. *+sMBR* row presents the performance with an additional sMBR training. It can be observed that E_s and E_{ph} based trainings outperform E_f

Table 2. Eval set WER on AMI, M-DE and M-FR corpora, and PER on TIMIT.

Error function →		E_f	E_s	E_{ph}
AMI		32.4	30.5	30.4
	+sMBR	30.4	28.4	28.4
M-DE		20.5	19.9	19.6
	+sMBR	19.7	18.7	19.0
M-FR		21.8	20.8	20.4
	+sMBR	20.6	18.9	19.0
TIMIT		22.3	21.2	21.3

based training. It is interesting to observe that, across all the three data sets, E_s or E_{ph} based trainings yield performances comparable to the E_f based training followed by sMBR.

4. ANALYSIS

This section presents an analysis of the proposed approach.

4.1. Generalisation to different architectures and front-ends

The proposed segment-level training approach does not presume any particular feature, front-end processing or ANN architecture. Nevertheless, a question that arises is whether the observations made in the previous section generalise across different architectures and front-ends. To investigate this, we conducted two ASR studies:

1. Training systems on the AMI data set with feature-space maximum likelihood linear regression (fMLLR) speaker transform based features and 25-frame splicing, concatenated with speaker-level iVectors, modelled with DNNs comprising six hidden layers with 2048 units each, and trained with dropout on speed-perturbed data. Table 3 presents the performance with the three error functions and with sMBR, as done before, in terms of WER.

Table 3. Performance on AMI data set with fMLLR+iVector front-end.

Error function →		E_f	E_s	E_{ph}
AMI		27.3	26.0	26.4
	+sMBR	25.1	23.9	24.1

2. Training convolutional neural network (CNN) based systems that take raw speech as input [25] on the M-DE data set. The CNN-based systems comprised four convolutional layers followed by three fully connected hidden layers with 1024 units each. Table 4 presents their performances in terms of WER.

Table 4. CNN-based system performance on M-DE data set.

Error function →	E_f	E_s	E_{ph}
M-DE	20.8	19.6	19.3

In both the studies, E_s and E_{ph} based trainings of the ANNs consistently yield better systems than E_f based training.

4.2. Effect of the segment duration normalisation

Different phones can have different durations; this can vary due to reasons such as the type of speech or speakers, for e.g. read versus spontaneous speech, native versus non-native speakers, etc. Also, the lengths of the silence portions can vary, for instance due to variations in a preceding voice activity detector's performance. Such differences in the durations of segments could affect the ANN training. Error functions E_s and E_{ph} inherently normalise the durations of the segments, and thus may handle their variations better.

To investigate this, we simulated a study on the TIMIT corpus, where silence was artificially added at the beginning and the end of each utterance. We considered two cases: (a) two seconds of silence is added at both the ends (4*s*/utt) and (b) five seconds silence is added at both the ends (10*s*/utt). We trained three hidden layer DNNs corresponding to E_f, E_s and E_{ph} error functions, as done earlier. Table 5 presents the results in terms of PER, when tested on silence-added utterances. It can be observed that, when trained with E_f, the phone recognition performance drastically degrades as the silence length increases at both the ends of the utterances, while when trained with E_s or E_{ph} the drop in the performance is significantly less. Investigating the ability to handle phone duration variations is part of our future work.

Table 5. PER on TIMIT corpus for the effect of segment duration normalisation study. 4*s*/utt and 10*s*/utt denote the addition of 2 seconds silence and 5 seconds of silence respectively at both the ends of the utterance.

Error function →		E_f	E_s	E_{ph}
TIMIT	4*s*/utt	23.0	22.0	21.8
	10*s*/utt	35.6	22.7	22.5

5. CONCLUSIONS AND FUTURE WORK

This paper established a link between local posterior-based confidence estimation in the acceptor HMM framework and the training of ANNs in hybrid HMM/ANN based systems. Through this link, we showed that the ANNs can be trained with error functions based on linguistic segments, such as sub-phonemic segments and phone segments as opposed to using a frame level cross entropy criterion. Through experimental studies on multiple corpora, we showed that such segment level trainings of ANNs yield better ASR systems. These gains in the performances are also sustained with sequence discriminative training. Furthermore, we demonstrated that the proposed segment level training approach (a) is generalisable across model architectures and front-ends, and (b) leads to systems that are robust to duration variations.

In addition to providing a link to acceptor HMMs, the proposed approach also provides a link to the Kullback-Leibler divergence based HMM (KL-HMM) framework [26, 27, 1]. More precisely, the sequence of target categorical distributions $(\mathbf{y}_{s_j})_{j=1}^{J}$ can be regarded as the parameters of the KL-HMM states. This allows us to incorporate the segment level training of ANNs into the embedded Viterbi training of KL-HMMs, where the parameters of the ANN and the KL-HMM are estimated in a recursive manner, and decoding is performed with local score based on KL-divergence. In other words, this link leads to a fully posterior based approach, where the ANN is trained with soft targets. Our future work will focus on establishing this link, along with the training of ANNs with error functions based on segments larger than phones, such as at word and utterance levels.

6. REFERENCES

[1] R. Rasipuram and M. Magimai.-Doss, "Acoustic and lexical resource constrained ASR using language-independent acoustic model and language-dependent probabilistic lexical model," *Speech Communication*, vol. 68, p. 23–40, Apr. 2015.

[2] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer Science & Business Media, 1994.

[3] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[4] G. Hinton *et al.*, "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.

[5] J. S. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Proceedings of Advances in Neural Information Processing Systems*, 1990, pp. 211–217.

[6] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[7] S. Austin, J. Makhoul, R. Schwartz, and G. Zavaliagkos, "Continuous speech recognition using segmental neural nets," BBN systems and technologies corp., Tech. Rep., 1991.

[8] O. Abdel-Hamid, L. Deng, D. Yu, and H. Jiang, "Deep segmental neural networks for speech recognition," in *Proceedings of Interspeech*, 2013, pp. 1849–1853.

[9] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2009, pp. 152–157.

[10] L. Kong, C. Dyer, and N. A. Smith, "Segmental recurrent neural networks," in *Proceedings of International conference on learning representations (ICLR)*, 2016.

[11] E. Beck, M. Hannemann, P. Dötsch, R. Schlüter, and H. Ney, "Segmental encoder-decoder models for large vocabulary automatic speech recognition," in *Proceedings of Interspeech*, 2018, pp. 766–770.

[12] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of Interspeech*, 2013, pp. 2345–2349.

[13] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proceedings of Interspeech*, 2016, pp. 2751–2755.

[14] G. Bernardis and H. Bourlard, "Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems," in *Proceedings of International conference on Spoken Language Processing*, vol. 3, 1998, pp. 775–778.

[15] G. Williams and S. Renals, "Confidence measures from local posterior probability estimates," *Computer Speech & Language*, vol. 13, no. 4, pp. 395–411, 1999.

[16] ——, "Confidence measures derived from an acceptor HMM," in *Proceedings of ICSLP*, no. 0644, 1998.

[17] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Transactions on Information Theory*, vol. IT-20, no. 4, pp. 405–417, 1974.

[18] S. Eguchi and J. Copas, "Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma," *Journal of Multivariate Analysis*, vol. 97, no. 9, 2006.

[19] R. Ullmann, R. Rasipuram, M. Magimai.-Doss, and H. Bourlard, "Objective intelligibility assessment of text-to-speech systems through utterance verification," in *Proceedings of Interspeech*, 2015.

[20] R. Rasipuram, M. Cernak, and M. Magimai.-Doss, "HMM-based non-native accent assessment using posterior features," in *Proceedings of Interspeech*, 2016.

[21] J. Makhoul, "Pattern recognition properties of neural networks," in *Proceedings of IEEE conference on Neural Networks for Signal Processing*, 1991, pp. 173–187.

[22] D. Imseng *et al.*, "Mediaparl: Bilingual mixed language accented speech database," in *Proceedings of Workshop on Spoken Language Technology (SLT)*, Dec 2012, pp. 263–268.

[23] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proceedings of International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[25] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proceedings of Interspeech*, 2013, pp. 1766–1770.

[26] G. Aradilla, J. Vepa, and H. Bourlard, "An acoustic model based on Kullback-Leibler divergence for posterior features," in *Proceedings of ICASSP*, 2007.

[27] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proceedings of Interspeech*, 2008.