MODELLING AUXILIARY FEATURES in TANDEM SYSTEMS

Mathew Magimai.-Doss[†], *Todd A. Stephenson*[‡], *Shajith Ikbal*[†], *Hervé Bourlard*[†]

[†]Dalle Molle Institute for Artificial Intelligence, CH-1920, Martigny, Switzerland
[†]Swiss Federal Institute of Technology (EPFL), CH-1015, Lausanne, Switzerland
[‡]Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

Tandem systems transform the cepstral features into posterior probabilities of subword units using artificial neural networks (ANNs), which are processed to form input features for conventional speech recognition systems. They have been shown to perform better than conventional speech recognition systems using cepstral features. Recent studies have shown that modelling cepstral features with auxiliary sources of knowledge leads to improvement in the performance of speech recognition systems. In this paper, we study two approaches to incorporate auxiliary knowledge sources such as pitch frequency, short-term energy, etc. (referred to as auxiliary features), in a tandem-based automatic speech recognition system. In the first approach, we model the auxiliary features in the process of training an ANN, which is later used to extract tandem-features. In the second approach, we extract the tandem-features from an ANN trained with cepstral features only and then model them jointly with auxiliary features. Recognition studies conducted on a connected word recognition task under clean and noisy conditions show that the performance of the tandem system can be improved by incorporating auxiliary features.

1. Introduction

The goal of the automatic speech recognition (ASR) systems is to produce a word transcription that best matches an acoustic sequence $X = x_1, \cdots, x_n, \cdots, x_N$, where N is the number of time frames. In state-of-the-art hidden Markov model (HMM) based ASR systems this problem is then formulated as modelling $p(Q, X)$, the evolution of the hidden state space $Q = \{q_1, \cdots, q_n, \cdots q_N\}$ and the observed space X [1]:

$$p(Q, X) \approx \prod_{n=1}^{N} p(x_n|q_n) \cdot P(q_n|q_{n-1}), \qquad (1)$$

where $q_n \in \{1, \cdots, k, \cdots, K\}$ and K is number of states.

In recent studies, it has been proposed that modelling the evolution of auxiliary information $A = \{a_1, \cdots, a_n, \cdots a_N\}$ along with Q and X (i.e. $p(Q, X, A)$ instead of $p(Q, X)$) could improve the performance of ASR [2, 3, 4, 5]:

$$p(Q, X, A) \approx \prod_{n=1}^{N} p(x_n, a_n|q_n) \cdot P(q_n|q_{n-1}) \qquad (2)$$

The implementation of such a system is straightforward, however this approach also implicitly models the dependency between the state q_n and the auxiliary feature a_n, which may be noisy. In such a case, it would be better to relax the joint distribution in (2) by assuming independence between a_n and q_n, yielding:

$$p(Q, X, A) \approx \prod_{n=1}^{N} p(x_n|q_n, a_n) \cdot p(a_n) \cdot P(q_n|q_{n-1}) \qquad (3)$$

Auxiliary features that were primarily investigated in the past such as pitch frequency, short-term energy, rate-of-speech (ROS) etc. were obtained directly from the speech signal [4]. This approach has been studied in the frameworks of both GMM-based HMM systems and hybrid HMM/ANN systems [4].

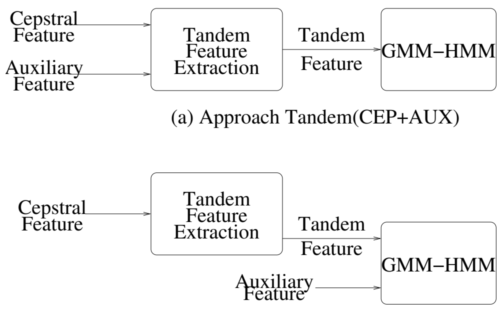
Traditional ASR systems use features such as Mel frequency cepstral coefficients (MFCCs), or perceptual linear prediction (PLP) features [6] etc., derived from the smoothed spectral envelope of the speech signal as the observation x_n. More recently, tandem systems have been proposed where the cepstral features are transformed into posterior probabilities using an ANN [7]. These posterior probabilites are then processed and fed as the input feature (tandem-feature) for a standard GMM-based ASR system. This has been shown to perform better than the state-of-the-art GMM-based ASR using cepstral features [7, 8].

In this paper, two different approaches to incorporate auxiliary features in a tandem system are investigated. In the first approach (Tandem(CEP+AUX)), hybrid HMM/ANN systems jointly modelling both the cepstral features and the auxiliary features based on (2) and (3) are trained [5]. The tandem-features are then extracted from the trained ANN of these systems. In the second approach (Tandem(CEP)+AUX), the tandem-features are extracted from the ANN of the hybrid HMM/ANN baseline system based on (1) and are modelled jointly with the auxiliary features in the framework of dynamic Bayesian networks (DBNs) as done in [4]. Figure 1 gives an illustration of the two approaches. We have studied this on the OGI Numbers database [9]. In both these approaches, significant improvement is achieved over ASR using cepstral features (in our case PLP cepstral coefficients). Moreover, the experimental studies also show that the performance of tandem systems could be further improved by incorporating auxiliary features.

The paper is organized in the following way. Sections 2 and 3 give a brief introduction about modelling auxiliary features and tandem systems, respectively. Section 4 describes the experimental setup and Section 5 presents the experimental studies. Finally, Section 6 summarizes our work with some conclusions.

2. Modelling Auxiliary Features

It can be observed from (2) and (3) that auxiliary features could be incorporated in standard ASR in different ways, such as ap-



(a) Approach Tandem(CEP+AUX)

(b) Approach Tandem(CEP)+AUX

Figure 1: *Illustraion of the two approaches to incorporate auxiliary features in tandem systems*

pending it to the feature vector as in (2) or conditioning the emission distribution as in (3). Although, implementation of a system based on (2) is easy, it is not obvious in the case of (3). In GMM-HMM systems this is realized through conditional Gaussians [3, 10]. In a hybrid HMM/ANN system, this can be realized by quantizing the auxiliary features and training an ANN corresponding to each of its discrete values [5], like gender modelling.

It is often observed that the estimation of the auxiliary features may not always be reliable, for example, the estimation of pitch frequency or short-term energy is error prone in noisy conditions. In such cases, it may be good to observe a_n during training and hide it, i.e. integrate over all possible values during recognition [4]. Refer to [4, 10] for further details about our approach of modelling auxiliary features in ASR.

3. Tandem Systems

In literature, GMM-HMM-based ASR systems and hybrid HMM/ANN-based ASR systems have been widely studied [1, 11]. GMM-HMM models are trained to maximize the likelihood of the data X, whereas an HMM/ANN model is trained to discriminate between the states so as to yield the posterior probability of state q_n.

A tandem system combines the discriminative feature of an ANN with Gaussian mixture modelling by using the processed posterior probabilities as the input feature for the GMM-HMM-based systems. This approach has been shown to yield significant improvement over conventional GMM-based ASR system using cepstral features in both clean and noisy conditions [7]. This approach has certain advantages such as, (a) it may allow us to make better use of the different probabilistic basis of the two systems and approaches developed for them. (b) it provides a framework where data from different databases could be used together, for instance, the ANN could be trained on any database [7].

The tandem system is trained in the following manner [7].

1. A hybrid HMM/ANN system is trained with task-independent or task-dependent data [7]. In our studies, it is the task-dependent data.
2. The task for which we have to train an ASR system, the training data of it is passed through the ANN to get the the posterior probabilities. In our case, it is the same data which is used to train hybrid HMM/ANN system.
3. Since the posterior probabilities obtained from the output of the ANN are very skewed, their logs are taken. This is similar to taking the value of the output units prior to the nonlinearity.
4. Principal component analysis (PCA) is performed on the features obtained in the previous step. The features are then decorrelated by projecting them along the eigenvectors. We refer to the resulting features as tandem-features. GMM-HMM-based ASR is then trained with the tandem-features.

During recognition, the test data is passed through the ANN and the log posterior proabilities are decorrelated by Karhunen-Loeve-transform (KLT) using the PCA statistics collected during training to obtain the tandem-features. The tandem-features are then fed to the trained HMMs and decoding is performed.

4. Experimental Setup

For our studies, we use the OGI-Numbers database which contains spontaneously spoken free-format numbers over telephone channel [9]. The definition of the training set, validation set and test set is similar to the one defined in [12]. The training set contains 3233 utterances (approximately 1.5 hours) and the validation set contains 357 utterances (used during ANN training). The test set consists of 1206 utterances.

We perform recognition experiments upon clean data. We also test our systems on versions with added noise using the Noisex-92 database [13]. We have studied it for factory (FACT) and lynx (LYNX) noise conditions at signal-to-noise ratios (SNR) 6dB and 12dB. The lexicon contains 30 different words.

4.1. Cepstral feature and auxiliary features used

We use 39 dimension PLP cepstral features comprised of 13 dimension PLP cepstral, along with their first and second order derivatives. The frame shift and frame size are 12.5 ms and 25 ms, respectively. Similar to earlier studies [4], we use the following auxiliary features: (a) Pitch (P), estimated using simple inverse filter tracking approach [14] with a 5-point median smoothing. The pitch estimator is reliable, and an evaluation of it can be found in [5]; (b) Short-term energy (E), estimated as the logarithm of the squared samples in a windowed frame; and (c) ROS (R), estimated called the *mrate* algorithm which was developed at ICSI and measures the ROS directly from the speech signal [15].

4.2. Modelling auxiliary features in tandem systems

We train different hybrid HMM/ANN systems. The input to the ANN are features at time frame n with 4 frame left and right context (9x39 vector) and the output is the posterior probabilities of 24 context-independent phonemes.

1. "System *B*": Hybrid HMM/ANN baseline system based on (1).
2. "System *H-A*": A hybrid HMM/ANN based on (2) for each auxiliary feature pitch, short-term energy, and ROS. The input to the ANN is a PLP feature vector appended with auxiliary feature at time frame n with 4 frame left and right context (9x40 vector).
3. "System *H-C*": As discussed earlier, in Sections 1 and 2, for hybrid HMM/ANN systems based on (3), the auxiliary feature has to be quantized [4, 5]. Similar to [4, 5], we quantize each type of auxiliary feature into three regions. We train an ANN for each of the regions by finding the nearest discrete region corresponding to the value of the auxiliary feature at that time frame.

With this setup, we can model auxiliary features in tandem system in the following ways

1. "Tandem(CEP+AUX)": Modelling the cepstral features and the auxiliary features in the framework of hybrid HMM/ANN (systems *H-A* and *H-C*) as in [4, 5]. Then, using these trained ANNs to extract tandem-features.
2. "Tandem(CEP)+AUX": Extracting the tandem-features from the hybrid HMM/ANN baseline system (system *B*), and then model them jointly with auxiliary features as done in [4, 10].

5. Experimental Studies

5.1. Tandem(CEP+AUX)

For the first approach, we use the HTK-toolkit [16] to train the GMM-HMM system with 80 context-dependent phonemes, 3 emitting states per phoneme and 12 mixtures per state. We train

1. *PLP*: A system with PLP features.
2. *Tandem(CEP)*: A system with tandem-features extracted from baseline HMM/ANN hybrid system (system *B*).
3. *Tandem(CEP+AUX-A)*: Three systems corresponding to the auxiliary features pitch, short-term energy and ROS with tandem-features extracted from their respective ANN of system *H-A*.
4. *Tandem(CEP+AUX-C)*: Three systems with tandem-features extracted from ANNs of system *H-C* corresponding to the auxiliary feature pitch, short-term energy and ROS.

In systems *Tandem(CEP+AUX-A)* and *Tandem(CEP+AUX-C)*, the auxiliary feature is always observed. In the case of system *Tandem(CEP+AUX-C)*, it means at any time frame n the ANN corresponding to the discrete value of the auxiliary feature a_n is used to extract tandem-features.

The results of the recognition studies are given in Table 1. It can be seen that the tandem systems perform better than the baseline system using PLP features in clean and noisy conditions. Comparing the tandem systems, in clean condition system *Tandem(CEP+AUX-A)* for auxiliary feature ROS performs better than the system *Tandem(CEP)*. The performance of system *Tandem(CEP+AUX-A)* for auxiliary features short-term energy and ROS degrades significantly in noisy conditions. The main reason for this is that the estimation of auxiliary features is not reliable. One solution would be to hide the continuous valued a_n; but it is not obvious how this could be done in the case of hybrid HMM/ANN systems .

5.2. Tandem(CEP)+AUX

For the second approach, we use DBNs [2, 10, 17]. DBNs like HMMs model $p(Q, X)$ or $p(Q, X, A)$ which puts them into the same family of models; but DBNs provide a more flexible framework for investigating the addition of variables to the modelling, the addition and deletion of statistical dependencies between component variables, and the hiding of some of the variables. We use the DBNs software developed in [10] to train ASR systems with 80 context-dependent phonemes, 3 emitting states per phoneme and 12 mixtures per state. We train following systems with DBNs

1. *PLP*: A baseline system using PLP features.
2. *Tandem(CEP)*: A tandem baseline system with tandem-features extracted from system *B*.

Table 1: Results of Tandem(CEP+AUX) approach where the tandem-features are extracted from hybrid HMM/ANN system modelling PLP features and auxiliary features. Results are reported for clean data (SNR=∞), SNRs of 6dB and 12dB. The performance is measured in-terms of word error rate (expressed in %). The best system for each condition is marked in boldface. Notations: *P*-Pitch, *E*-Short-term energy, *R*-ROS

		LYNX		FACT	
	∞	12	6	12	6
PLP	7.3	11.6	20.0	16.2	37.6
Tandem(CEP)	5.1	9.4	**16.2**	**13.2**	**25.6**
Tandem(CEP+AUX-A) (*P*)	5.1	**9.1**	16.3	13.8	26.2
Tandem(CEP+AUX-C) (*P*)	5.5	9.9	16.4	14.6	31.2
Tandem(CEP+AUX-A) (*E*)	5.7	19.3	46.6	34.0	70.1
Tandem(CEP+AUX-C) (*E*)	5.7	10.9	19.3	15.8	28.8
Tandem(CEP+AUX-A) (*R*)	**4.8**	15.0	34.6	26.3	59.0
Tandem(CEP+AUX-C) (*R*)	6.0	10.7	18.1	15.9	30.8

3. *Tandem(CEP)+AUX-A*: Three systems corresponding to the different auxiliary features pitch, short-term energy and ROS based on (2). Here, the tandem-features are augmented with the auxiliary feature (assuming $x_n \perp\!\!\!\perp a_n \mid q_n$).
4. *Tandem(CEP)+AUX-C*: Three systems corresponding to the different auxiliary feature pitch, short-term energy ROS based on (3). Here, the auxiliary feature conditions the emission distribution.

The auxiliary feature is observed throughout the training. During recognition, we also performed experiments hiding the auxiliary feature [4, 10].

The results of the recognition studies are given in Table 2. The tandem systems again perform better than the PLP baseline system in both clean and noisy conditions. When comparing between tandem systems in clean condition the system *Tandem(CEP)+AUX-A* performs better than the system *Tandem(CEP)*. In order to verify that this improvement is not due to an increase in the number of parameters, we trained a *Tandem(CEP)* system with 18 mixtures. The performance of this system is 5.1% in clean, 8.8% (LYNX SNR 12dB), 15.4% (LYNX SNR 6dB), 12.3% (FACT SNR 12dB) and 24.6% (FACT SNR 6dB). *Tandem(CEP)+AUX-A* performs better than this system in all conditions when the auxiliary features short-term energy and ROS are hidden.

6. Summary and Conclusion

In this paper, we studied two approaches to incorporate auxiliary features in tandem-based ASR system. In the first approach, we model cepstral features and auxiliary features by ANNs which are later used to extract tandem-features. In the second approach, we extract tandem-features through an ANN trained on cepstral features only and then, model them with auxiliary features. Experiments conducted in both clean and noisy conditions shows that the tandem system performs better than conventional GMM-HMM systems using cepstral features. Our studies also show that the performance of the tandem system could be enhanced further, especially by modelling tandem-features jointly with auxiliary features.

In earlier studies, tandem systems have been shown to perform better than the conventional systems in noisy conditions;

Table 2: Results of Tandem(CEP)+AUX approach where the tandem-features are extracted from a hybrid HMM/ANN baseline system and, are modelled along with auxiliary features using DBNs. For systems using auxiliary features, the first row corresponds to the case when the auxiliary features are observed and the second row to the case when the auxiliary features are hidden. Results are reported for clean data (SNR=∞), SNRs of 6dB and 12dB. The performance is measured in-terms of word error rate (expressed in %). † Systems performing signficantly better than *Tandem(CEP)* system (with 95% confidence). The best system(s) for each condition is marked boldface. Notations: *P*-Pitch, *E*-Short-term energy, *R*-ROS

		LYNX		FACT	
	∞	12	6	12	6
PLP	7.3	16.3	33.3	24.6	46.9
Tandem(CEP)	5.2	9.3	15.4	13.0	24.6
Tandem(CEP)+AUX-A (*P*)	4.9	8.3†	14.6	12.5	25.0
	4.9	8.6	15.3	12.3	24.7
Tandem(CEP)+AUX-C (*P*)	5.4	9.6	15.9	13.4	25.5
	5.8	9.6	16.1	13.0	24.7
Tandem(CEP)+AUX-A (*E*)	4.9	8.4	14.8	12.6	24.2
	4.8	**8.2**†	15.0	**11.9**†	**23.7**
Tandem(CEP)+AUX-C (*E*)	6.1	10.6	17.7	13.8	25.5
	5.5	9.6	17.0	13.6	24.8
Tandem(CEP)+AUX-A (*R*)	**4.7**	**8.2**†	14.8	12.8	26.6
	4.8	**8.2**†	**14.3**	12.3	24.2
Tandem(CEP)+AUX-C (*R*)	5.7	10.0	16.7	13.7	25.6
	5.6	9.8	16.3	13.4	25.4

but when trained on speech of multiple conditions [7]. In our studies, we observe that the tandem system trained only with clean speech can perform better than conventional systems in noisy conditions.

We performed recognition studies on systems *B*, *H-A* and *H-C* with 24 context-independent phonemes in clean speech condition. The system *H-C* performs better than system *B* (9.6%) for pitch (8.4%) and short-term energy (8.2%); but system *Tandem(CEP+AUX-C)* performs worse than the system *Tandem(CEP)* for pitch and short-term energy (see Table 1). Similar trend has been observed earlier in literature[8], where the improvements in the context-independent system does not shows up in the context-dependent GMM-HMM system using tandem features. In our case, the reason for this could be the switching between the ANNs corresponding to the discrete-valued auxiliary feature, as this may be affecting the PCA analysis part of the tandem-feature extraction. This has to be further investigated.

7. Acknowledgements

This work was supported by the Swiss National Science Foundation (NSF) under grant MULTI (2000-068231.02/1) and Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2. The NCCR is managed by the Swiss NSF on behalf of the federal authorities. This paper has benefitted from the valuable comments and suggestions of Prof. Hynek Hermansky, Sunil Sivadas and Guillaume Lathoud. The authors would like to thank Sunil Sivadas for his help in setting up the HTK system and drSpeech tools used for PCA analysis. We would also like to thank Joanne Moore for proofreading this paper.

8. References

[1] L. R. Rabiner and H. W. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs New Jersey, 1993.

[2] Geoffrey G. Zweig, *Speech Recognition with Dynamic Bayesian Networks*, PhD dissertation, University of California, Berkeley, 1998.

[3] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, “Multiple-regression hidden Markov model,” in *ICASSP*, 2001, pp. 513–516.

[4] T. A. Stephenson, M. Magimai.-Doss, and H. Bourlard, “Speech recognition with auxiliary information,” *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 189–203, May 2004.

[5] M. Magimai.-Doss, T. A. Stephenson, and H. Bourlard, “Using pitch frequency information in speech recognition,” in *Eurospeech*, September 2003, pp. 2525–2528.

[6] H. Hermansky, “Perceptual linear predictive(PLP) analysis of speech,” *JASA*, vol. 87, no. 4, pp. 1738–1752, 1990.

[7] H. Hermansky, D. Ellis, and S. Sharma, “Tandem connectionist feature stream extraction for conventional hmm systems,” in *ICASSP*, 2000, pp. III–1635–1638.

[8] D. P. W. Ellis, R. Singh, and S. Sivadas, “Tandem acoustic modeling in large-vocabulary recognition,” in *ICASSP*, 2001.

[9] R. A. Cole, M. Fanty, M. Noel, and T. Lander, “Telephone speech corpus development at CSLU,” in *ICLSP*, September 1994.

[10] Todd A. Stephenson, *Speech recognition with auxiliary information*, PhD dissertation, Swiss Federal Institutue of Technology (EPFL), Lausanne, 2003.

[11] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[12] N. Mirghafori and N. Morgan, “Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers,” in *ICLSP*, 1998, pp. 743–746.

[13] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, “The NOISEX-92 study on the effect of additive noise on automatic speech recognition,” Tech. Rep., DRA Speech Research Unit, Malvern, England, 1992.

[14] J. D. Markel, “The SIFT algorithm for fundamental frequency estimation,” *IEEE Trans. Audio and Electroacoustics*, vol. 20, pp. 367–377, 1972.

[15] Nelson Morgan and Eric Fosler-Luisser, “Combining multiple esitmators of speaking rate,” in *ICASSP*, Seattle, 1998, pp. 729–732.

[16] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “Hidden Markov model toolkit V2.1 reference manual,” Technical report, Speech group, Engineering Department, Cambridge University, UK, March 1997.

[17] Kevin Patrick Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, University of California, Berkeley, 2002.