IDIAP RESEARCH REPORT

Modelling auxiliary information (pitch frequency) in hybrid HMM/ANN based ASR systems

Mathew Magimai.-Doss [a] [b]
Todd A. Stephenson [a] [b] Hervé Bourlard [a] [b]

IDIAP–RR 02-62

December 2002

Institut Dalle Molle
d'Intelligence Artificielle
Perceptive • CP 592 •
Martigny • Valais • Suisse

téléphone +41–27–721 77 11
télécopieur +41–27–721 77 12
adr.él. secretariat@idiap.ch
internet http://www.idiap.ch

[a] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), CH-1920, Martigny, Switzerland

[b] The Swiss Federal Institute of Technology (EPFL), CH-1015, Lausanne, Switzerland

Rapport de recherche de l'IDIAP 02-62

Modelling auxiliary information (pitch frequency) in hybrid HMM/ANN based ASR systems

Mathew Magimai.-Doss Todd A. Stephenson Hervé Bourlard

December 2002

Résumé. Automatic Speech Recognition (ASR) systems typically use smoothed spectral features as acoustic observations. In recent studies, it has been shown that complementing these standard features with auxiliary information could improve the performance of the system. The previously proposed systems have been studied in the framework of GMMs. In this paper, we study and compare different ways to include auxiliary information in state-of-the-art hybrid HMM/ANN system. In the present paper, we have focused on pitch frequency as the auxiliary information. We have evaluated the proposed system on two different ASR tasks, namely, isolated word recognition and connected word recognition. Our results complement the previous efforts to incorporate auxiliary information in ASR system and also show that pitch frequency can indeed be used in ASR systems to improve the recognition performance.

1 Introduction

In standard automatic speech recognition systems, at each time frame n, hidden Markov model (HMM) estimates the likelihood (also called emission probability) of the acoustic observation x_n being produced, given the hidden state q_n [RJ93]

$$p(x_n|q_n) \tag{1}$$

where $q_n \in \{1,\cdots,k,\cdots,K\}$. This is typically estimated using Gaussian Mixture Models (GMMs) or Artificial Neural Network (ANN).

Along with the approach taken to model the emission distribution, the choice of acoustic features has direct impact on the performance of ASR. Standard ASR systems use Mel Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Prediction (PLP) cepstral coefficients as acoustic features. In recent studies, it has been shown that these standard acoustic features can be supplemented with additional information called auxiliary information a_n, to improve the performance of the ASR system [FNSS01, SEMDB02]. The auxiliary information is a secondary information which may not be directly useful for ASR, e.g., gender information. The auxiliary information can be a knowledge such as gender information, or extra features extracted from the speech signal, such as articulatory features, pitch frequency, rate-of-speech etc. Modelling the auxiliary information in standard ASR may help to make the ASR system more robust to speaker variability. In [Sie95] for instance, two acoustic models corresponding to fast and slow speech were trained to compensate the effects of speaking rate.

The auxiliary information can be incorporated in a standard ASR in different ways that will be studied here, including:

(a) Augmenting the standard features with the auxiliary information and estimating the emission distribution using the augmented features.

$$p(x_n,a_n|q_n) \tag{2}$$

(b) Conditioning the emission distribution with a_n.

$$p(x_n|q_n,a_n) \tag{3}$$

A typical example of such a system is gender modelling where the a_n is discrete valued, $a_n \in \{Male, Female\}$ [KMC91].

While implementation of (2) seems easy; implementation of a system based upon (3) is not straightforward, particularly when a_n is continuous valued. Approaches to realize such systems when the emission distribution is modelled by GMMs were recently proposed in [FNSS01, SEMDB02].

In this paper, we study different ways in which the auxiliary information can be introduced in a hybrid HMM/ANN based ASR. Hybrid HMM/ANN system naturally address both the time-dependence and the within feature vector dependence assumption. Typically, the ANN is used as a classifier. There are known advantages in using an ANN to model emission distribution such as modelling discrimination, modelling higher-order correlation between the components of the feature vector, access to posteriors etc [BM94]. In hybrid HMM/ANN systems, the emission probability is estimated from the state posterior distribution (which is discrete) obtained from the output of the ANN, whereas, in HMM/GMM systems the emission probability is estimated from the mixture of Gaussian distributions (which is continuous). Hence, there is no direct extension to the approach suggested in [FNSS01, SEMDB02]. Also, in [SEMDB02, SMDB02a] it has been shown that observing the auxiliary information during training and hiding it during recognition may help in improving the performance of the system. As we will see in the next section, this is not always possible in case of hybrid HMM/ANN system.

In Section 2, we present the different approaches to model auxiliary information in a hybrid HMM/ANN based ASR. Section 3 then describes our system and the experimental studies, before concluding with an analysis of the results obtained.

2 Introducing Auxiliary Information

Standard HMM based ASR models $p(Q,X)$ [RJ93], the evolution of the observed space $X = \{x_1,\cdots,x_n,\cdots,x_N\}$ and the hidden state space $Q = \{q_1,\cdots,q_n,\cdots,q_N\}$ for time $n = 1,\cdots,N$ as:

$$p(Q,X) \approx \prod_{n=1}^{N} p(x_n|q_n) \cdot P(q_n|q_{n-1}) \tag{4}$$

In case of hybrid HMM/ANN based ASR $p(x_n|q_n)$ is replaced by the scaled likelihood $p_{sl}(x_n|q_n)$, which is estimated as [BM94]:

$$p_{sl}(x_n|q_n) = \frac{p(x_n|q_n)}{p(x_n)} = \frac{P(q_n|x_n)}{P(q_n)} \tag{5}$$

For incorporating auxiliary information $A = \{a_1,\cdots,a_n,\cdots,a_N\}$, we have to model $p(Q,X,A)$. The auxiliary information can be discrete valued i.e. $a_n \in \{1,\cdots,l,\cdots,L\}$ or continuous valued. The simplest and most common practice is to augment feature vector x_n with a_n yielding $y_n = (x_n,a_n)$ and modelling evolution of $Y = \{y_1,\cdots,y_n,\cdots y_N\}$ over the hidden state space Q similar to (4), resulting in:

$$\begin{aligned} p(Q,X,A) \quad &= p(Q,Y) \\ &\approx \prod_{n=1}^{N} p(y_n|q_n) \cdot P(q_n|q_{n-1}) && (6) \\ &\approx \prod_{n=1}^{N} p(x_n|q_n,a_n) \cdot p(a_n|q_n) \cdot P(q_n|q_{n-1}) && (7) \end{aligned}$$

The implementation of such a system is straight-forward, irrespective of whether the auxiliary information is discrete or continuous valued. As it can be observed from (7), this approach also implicitly models the dependency between the state q_n and the auxiliary information a_n, which may be noisy in some cases. For example, if the auxiliary information is gender then it cannot tell anything about the state q_n or what has been spoken. In such a case, it would be better to relax the joint distribution in (7) by assuming conditional independence between a_n and q_n, yielding

$$p(Q,X,A) \approx \prod_{n=1}^{N} p(x_n|q_n,a_n) \cdot P(a_n) \cdot P(q_n|q_{n-1}) \tag{8}$$

If the auxiliary information is discrete valued then, a system based upon (8) could be realized by training an ANN corresponding to each discrete value. This is similar to the case of gender modelling, where acoustic models for male and female speaker are simply trained separately. In case of continuous valued a_n, it is not evident how to implement a hybrid HMM/ANN system according to (8). For the case of emission distribution modelled by Gaussian such a system is realized using conditional Gaussian [LJ01, FNSS01, SEMDB02], where the first order moment of the distribution is a linear regression upon the auxiliary information. An alternative would be to relax (7) by assuming conditional independence between x_n and a_n i.e.

$$p(Q,X,A) \approx \prod_{n=1}^{N} p(x_n|q_n) \cdot p(a_n|q_n) \cdot P(q_n|q_{n-1}) \tag{9}$$

This would mean that x_n and a_n are two separate inputs. They are connected to the output layer through different hidden layers. Such a system is similar to the system in (6); except that the correlation between x_n and a_n will not be modelled.

The auxiliary information sometimes may not be available or the estimation may be noisy. This is the case, for e.g., with pitch frequency as well as speaking rate estimation, which is not always perfect. In such a case, it may be good to observe the auxiliary information during training and hide it i.e. integrate over all possible values during recognition [SEMDB02]. The auxiliary information then can be hidden in two ways depending upon how the auxiliary information is treated. The auxiliary information can be a static information such as gender information. In such a case, the discrete valued auxiliary information can be hidden in the following way:

$$p(Q,X) = \sum_{l=1}^{L} p(Q,X,A=l) \tag{10}$$

This would mean running the decoder over all the L different systems and summing their output. If the auxiliary information is a dynamic variable, the auxiliary information can be hidden by marginalizing the distribution $p(x_n,a_n|q_n)$ over a_n to obtain the emission distribution $p(x_n|q_n)$ and performing decoding according to (4) [SEMDB02, SMDB02a]. Again in hybrid HMM/ANN system it is not clear how it could be done when the auxiliary information is continuous valued. However, for the case of discrete valued auxiliary information, the auxiliary information could be hidden to estimate $p(x_n|q_n)$ in the following way:

$$\begin{aligned} p(x_n|q_n) = & \sum_{l=1}^{L} p(x_n,a_n=l|q_n) && (11) \\ \approx & \sum_{l=1}^{L} p(x_n|q_n,a_n=l) \cdot P(a_n=l) && (12) \end{aligned}$$

and performing decoding according to (4). Equation (12) corresponds to (8), when the auxiliary information is hidden.

3 Experiments

3.1 Systems

We study 4 different hybrid HMM/ANN systems.
System 1: Baseline system based on (4).
System 2: System with x_n and a_n based on (6); a_n is continuous valued.
System 3: System with a_n conditionally independent of q_n based on (8); a_n is discrete valued.
System 4: System with x_n conditionally independent of a_n based on (9); a_n is continuous valued.

3.2 Database and Features

The above systems are studied for two different tasks of ASR, isolated word recognition and connected word recognition. We use PhoneBook speech corpus for speaker-independent task-independent, small vocabulary (75 words) isolated word recognition [PFW$^+$95]. For connected word recognition task, we use OGI Numbers speech corpus which contains free-format numbers spontaneously spoken by different speakers [CFL94]. The definitions of the training, validation, and evaluation sets are similar to [DBD$^+$97] and [MM98], for PhoneBook corpus and OGI Numbers corpus, respectively.

There are 42 context-independent phones including silence, each modelled by a single emitting state in the systems trained on PhoneBook corpus. The acoustic vector x_n is the MFCCs extracted from the speech signal using a window of 25 ms with a shift of 8.3 ms. Cepstral mean subtraction and energy normalization are performed. Ten MFCCs, the first-order derivatives (delta) of the ten MFCCs and the c_0 (energy coefficient) is extracted for each time frame, resulting in 21 dimension acoustic vector.

In the systems trained on OGI Numbers, there are 27 context-independent phones including silence, each modelled by a single emitting state. The acoustic observation x_n consists of 12th order PLP plus the energy cepstral features, their deltas and their delta-deltas extracted from a 25 ms speech signal with a frame shift of 12.5 ms.

The auxiliary information used in our studies is pitch frequency, which is extracted using simple inverse filter tracking (SIFT) algorithm [Mar72]. This method retains the advantages of autocorrelation and cepstral analysis techniques. A 5-point median smoothing is performed on the pitch frequency contour. The pitch frequencies are normalized by the highest pitch frequency which is 400 Hz in our case (same for all utterances), before being used in the systems where a_n is continuous valued. The normalization is done in order to avoid saturation of the sigmoids [LBOM98]. Theoretically, this normalization should not affect the performance of the system.

3.3 Experimental Studies

In Section 2 when we described the hiding strategies, we observed that the auxiliary information can be static information or dynamic information. For example, existence or nonexistence of pitch frequency at the frame level conveys information about voicing; but at the same time the average of pitch frequency over an utterance can convey gender information. Similarly, rate-of-speech is more of a suprasegmental information than a segmental information.

We performed two different set of experiments to study the use of pitch frequency as auxiliary information. In the first set of experiments the pitch frequency is treated as static information and in the second set of experiments pitch frequency is treated as dynamic information.

3.3.1 Pitch frequency treated as static information

In this first set of experiments, we studied Systems 1 and 3 only. We use the average pitch frequency computed over an entire utterance as the auxiliary information; such information can be considered to be independent of the state q_n. Furthermore, we investigate the case where the auxiliary information is hidden according to (10), during recognition.

The PhoneBook database was used for this study. A baseline (System 1) MultiLayer Perceptron (MLP) was trained with the 21 dimension MFCC feature vector as the input with the left and right context of four frames each. To implement System 3, the average pitch frequency of each training set utterance is computed and then, the average pitch frequencies are quantized into two discrete regions. This approach can be compared to the gender modelling approach; except that both the genders will be present in the discrete regions as the male and female pitch frequencies overlap at higher and lower extremes, respectively. For each of these discrete regions, a multilayer perceptron (MLP) is trained with the same 21 dimension feature vectors used earlier to train the baseline system, with a left and right context of four frames each. During recognition, we have three options:

1. O: The auxiliary information is observed. This is done by computing the average pitch frequency of the test utterance and selecting the MLP corresponding to the nearest of the two discrete regions for decoding and decode the test utterance.
2. H: Auxiliary information is hidden according to (10).
3. M: Decoding is done parallely on the two systems, similar to the hidden case and the maximum output is picked for decision making (equivalent to replacing the *sum* operation by *max* operation in (10)). This is the common approach adopted during recognition in gender modelling.

The results of this study are given in Table 1. System 1 performs better than System 3 in all cases. The results obtained for System 3 show the advantage of hiding the auxiliary information over observing the auxiliary information, during recognition.

TAB. 1 – *Comparing the performance of different systems where pitch frequency is used as a static auxiliary information. The performance of the systems are expressed in terms of word error rate (WER).*

	Auxiliary Information	Performance
System 1	N.A	5.4%
System 3	O	12.3%
System 3	H	6.8%
System 3	M	7.1%

3.3.2 Pitch frequency treated as dynamic information

In this second set of experiments, we study all the systems described in Section 3.1. The pitch frequency estimated at each time frame is used as the auxiliary information in this set of experiments. The studies were conducted on both PhoneBook database and OGI Numbers database.

The PhoneBook systems were trained with the 21 dimension MFCC features. The OGI Numbers systems were trained with the 39 dimension PLP features. The baseline systems were trained with the standard acoustic features.

The System 2 was trained by concatenating the pitch frequency to the standard acoustic feature vector at every frame i.e. the input layer contains additional input corresponding to the auxiliary information.

The System 3 was implemented in the following manner.

1. The pitch frequency contour is computed for all the training utterances.
2. The pitch frequencies are then vector quantized into three discrete regions.
3. An MLP corresponding to each of the discrete regions is trained by finding the nearest discrete region corresponding to the value of the auxiliary information at that frame. The only exception is that the silence regions are observed by all the three MLPs.

During recognition, we study two strategies as we did in Section 3.3.1, namely, auxiliary information observed, (O) and auxiliary information hidden, (H). When the auxiliary information is observed, during decoding the output of the MLP corresponding to discrete region nearest to the auxiliary information observed at that frame is used for estimating the emission probability. When the auxiliary information is hidden the decoding is performed according to (12).

System 4 has similar architecture as the one of the baseline system except that the output layer has an additional input corresponding to the auxiliary information. In our present studies, we have used the pitch frequency at that frame as the auxiliary information. In future, we would like to model the time correlation across the auxiliary information by introducing a separate hidden layer for the auxiliary information.

The results of the studies conducted on PhoneBook database and OGI Numbers database are given in Table 2 and Table 3, respectively. In both the studies, System 2 performs better than all other systems.

4 Conclusion

In this paper, we studied how auxiliary information can be incorporated in state-of-the-art hybrid HMM/ANN system, which allows us to take the benefits of ANN. The results obtained complements the previous efforts to model auxiliary information within the frame work of HMM/GMM and dynamic Bayesian networks [FNSS01, SEMDB02].

Comparison between the performance of System 3 in Table 1 and Table 2 (for observed (O) and hidden (H) cases) shows that better acoustic models could be obtained, when pitch frequency is treated as dynamic auxiliary information.

TAB. 2 – *Comparison of the performance of the systems trained on PhoneBook database with pitch frequency as dynamic auxiliary information. The performance is expressed in-terms of WER.*

	Auxiliary Information	Performance
System 1	N.A	5.4%
System 2	O	4.2%
System 3	O	6.9%
System 3	H	5.1%
System 4	O	4.7%

TAB. 3 – *Comparison of the performance of the systems trained on OGI Numbers database with pitch frequency as dynamic auxiliary information. The performance is expressed in-terms of WER.*

	Auxiliary Information	Performance
System 1	N.A	12.1%
System 2	O	11.3%
System 3	O	11.5%
System 3	H	12.1%
System 4	O	13.1%

Comparison of the performance of Systems 2 and 3 against System 4 in Tables 2 and 3 show that standard acoustic feature and pitch frequency are correlated and modelling this correlation effectively can improve the performance of the system. The results also suggest that pitch frequency can be used as a component of feature vector.

In System 3, when auxiliary information is observed or hidden it is done throughout the test utterance. The results of System 3 in Tables 2 and 3 indicates that it is worth investigating when to hide the auxiliary information during recognition.

Though, our results show that the simple concatenation approach yields the best result;but, this may not be true for other auxiliary information [SMDB02b]. In future, we would like to model other auxiliary information such as rate-of-speech and short-time energy in the context of modelling speaker variability in spontaneous speech.

Acknowledgement

This work was supported by the Swiss National Science Foundation under grants PROMO (21-57245.99) and BN_ASR (20-64172.00). We would also like to thank TCTS, FPMS, Belgium for providing us the segmentation of PhoneBook database.

Références

[BM94] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition - A Hybrid Approach.* Kluwer Academic Publishers, 1994.

[CFL94] R. A. Cole, M. Fanty, and T. Lander. Telephone speech corpus at CSLU. In *Proceedings of Int. Conf. Spoken Language Processing*, Yokohama, Japan, September 1994.

[DBD^{+}97] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite. Hybrid HMM/ANN systems for training independent tasks: Experiments on 'PhoneBook' and related improvements. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 524–528, 1767-1770, 1997.

[FNSS01] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama. Multiple-regression hidden Markov model. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 513–516, 2001.

[KMC91] Yochai Konig, Nelson Morgan, and Claudia Chandra. GDNN: A gender-dependent neural network for continuous speech recognition. Technical Report TR-91-071, ICSI, Berkeley, Berkeley, Califronia, USA, December 1991.

[LBOM98] Yann LeCun, Leon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Effiecient BackProp. In Genevieve N. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, number 1524 in Lecture Notes in Computer Science, chapter 1, pages 9–50. Springer-Verlag, Berlin Heidelberg, 1998.

[LJ01] S. L. Lauritzen and F. Jensen. Stable local computations with conditional gaussian distributions. *Statistics and Computing*, 11(2):191–203, April 2001.

[Mar72] John D. Markel. The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. Audio and Electroacoustics*, 20:367–377, 1972.

[MM98] N. Mirghafori and N. Morgan. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. In *Proceedings of Int. Conf. Spoken Language Processing*, pages 743–746, Sydney, November 1998.

[PFW^{+}95] J. F. Pitrelli, C. Fong, S. H. Wong, J. R. Spitz, and H. C. Leung. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 1767–1770, May 1995.

[RJ93] L. R. Rabiner and H. W. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs New Jersey, 1993.

[SEMDB02] Todd A. Stephenson, Jaume Escofet, Mathew Magimai-Doss, and H. Bourlard. Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables. In *IEEE Conference on Neural Network Signal Processing*, Martigny, Switzerland, September 2002.

[Sie95] Matthew A. Siegler. *Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition.* MS dissertation, Carnegie Mellon University, Department of Electrical and Computer Engineering, December 1995.

[SMDB02a] Todd A. Stephenson, Mathew Magimai-Doss, and H. Bourlard. Auxiliary variables in conditional gaussian mixtures for automatic speech recognition. In *Proceedings of Int. Conf. Spoken Language Processing*, Denver, US, September 2002.

[SMDB02b] Todd A. Stephenson, Mathew Magimai-Doss, and Hervé Bourlard. Speech recognition of spontaneous, noisy speech using auxiliary information in Bayesian networks. IDIAP-RR 44, IDIAP, 2002.