Using pitch frequency information in speech recognition

Mathew Magimai.-Doss, Todd A. Stephenson, Hervé Bourlard

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), CH-1920, Martigny, Switzerland
The Swiss Federal Institute of Technology (EPFL), CH-1015, Lausanne, Switzerland
mathew@idiap.ch, todd@idiap.ch, bourlard@idiap.ch

Abstract

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 pitch frequency could improve the system performance of the system [1, 2]. While previously proposed systems have been studied in the framework of HMM/GMMs, in this paper we study and compare different ways to include pitch frequency in state-of-the-art hybrid HMM/ANN system. We have evaluated the proposed system on two different ASR tasks, namely, isolated word recognition and connected word recognition. Our results show that pitch frequency can indeed be used in ASR systems to improve the recognition performance.

1. Introduction

Speech is produced by a linear time-variant vocal tract system excited by the vibration of vocal cords. The acoustic speech signal mainly contains two kinds of information, namely, source information and vocal tract system information. Traditional ASR systems use features derived from the smoothed spectral envelope of the speech signal which basically represent the characteristics of the vocal tract system (alleviating the knowledge of voice source characteristics), e.g. perceptual linear prediction (PLP) features[3].

Voice source characteristic such as pitch is a perceptual quantity; but its acoustic correlate (rate of vibration of vocal cords) referred to as pitch frequency, can be estimated from the speech signal. Pitch frequency can convey different information, information about the speaker; its existence or non-existence can convey information about the type of sound (voiced or unvoiced); its variation across time can convey prosodic information. Hence, pitch frequency is not an ideal source of information for ASR. It has been observed in literature that pitch frequency affects the estimation of the spectral envelope, in particular, the estimation of the spectral peaks, making the standard acoustic features sensitive to changes in pitch frequency, e.g. [3]. Thus, we may expect certain correlation between standard acoustic feature and pitch frequency, for example [1] illustrates a negative correlation between 7th Mel cepstral coefficient and logarithm of the pitch frequency of a phoneme sample. In recent studies, it has been shown that the standard acoustic features can be supplemented with additional information such as pitch frequency to improve the performance of ASR system [1, 2].

In standard automatic speech recognition systems, at each time frame n, hidden Markov models (HMMs) estimate the likelihood (also called emission probability) of the acoustic observation x_n being emitted on a specific state q_n [4]

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

where $q_n \in \{1, \cdots, k, \cdots, K\}$, set of possible HMM states. This is typically estimated using Gaussian Mixture Models (GMMs) or Artificial Neural Network (ANN). In incorporating pitch frequency (F_{0n}) at time frame n, we can model (1) in the following ways:

(a) Augmenting the standard features with pitch frequency F_{0n} and estimating the emission distribution using the augmented features.

$$p(x_n, F_{0n}|q_n) \tag{2}$$

(b) Conditioning the emission distribution upon F_{0n}.

$$p(x_n|q_n, F_{0n}) \tag{3}$$

A particular example of such a system is gender modelling. In gender modelling [5], two different acoustic models are trained corresponding to each gender using their respective training data. During recognition, there are different options such as one can run a gender recognizer and pick the acoustic models accordingly or pick the one which gives the best match ($\max$ operation) for decision making or hide the gender information (integrating over all possible values).

While implementing (2) seems easy, the implementation of a system based upon (3) is not straightforward, if F_{0n} is continuous valued. Approaches to realize systems using (3) when the emission distribution is modelled by GMMs were recently proposed in [1, 2].

In this paper, we study different ways in which the pitch frequency information can be introduced in a hybrid HMM/ANN based ASR. Hybrid HMM/ANN systems naturally address both the time-dependence and the within feature vector dependence assumption. There are known advantages in using an ANN to model emission distribution such as better discrimination, modelling higher-order correlation between the components of the feature vector, access to posterior probabilities etc [4]. 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 [1, 2]. Also, in [2] it has been shown that observing the pitch frequency 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 pitch frequency in 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. Modelling Pitch Frequency in Hybrid HMM/ANN ASR

Standard HMM based ASR models $p(Q, X)$ [4], 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 [4]:

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

For incorporating pitch frequency information $F_0 = \{F_{01}, \cdots, F_{0n}, \cdots, F_{0N}\}$, we have to model $p(Q, X, F_0)$. The pitch frequency can be discrete valued i.e. $F_{0n} \in \{1, \cdots, l, \cdots, L\}$ or continuous valued. The simplest and most common practice is to augment the feature vector x_n with F_{0n} and model the evolution of the augmented feature vector over the hidden state space Q similar to (4), resulting in:

$$p(Q, X, F_0) \approx \prod_{n=1}^{N} p(x_n|q_n, F_{0n}) \cdot p(F_{0n}|q_n) \cdot P(q_n|q_{n-1}) \tag{6}$$

The implementation of such a system is straightforward, irrespective of whether the pitch frequency is discrete or continuous valued. As it can be observed from (6), this approach also implicitly models the dependency between the state q_n and the pitch frequency F_{0n}, which may be noisy. For example, pitch frequency 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 (6) by assuming independence between F_{0n} and q_n, yielding

$$p(Q, X, F_0) \approx \prod_{n=1}^{N} p(x_n|q_n, F_{0n}) \cdot P(F_{0n}) \cdot P(q_n|q_{n-1}) \tag{7}$$

If the pitch frequency is discrete valued then, a system based upon (7) 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 F_{0n}, it is not evident how to implement a hybrid HMM/ANN system according to (7). For the case of emission distribution modelled by Gaussian such a system is realized using conditional Gaussian [6, 1, 2], where the first order moment of the distribution is a linear regression upon the pitch frequency.

It has been shown in literature that pitch frequency estimation is error prone [7]. In such a case, it may be good to observe F_{0n} during training and hide it i.e. integrate over all possible values during recognition [2]. The pitch frequency then can be hidden in two ways depending upon how the pitch frequency is treated. The pitch frequency can be a static information (average pitch frequency over the entire utterance, e.g., gender modelling). In such a case, the discrete valued pitch frequency can be hidden in the following way:

$$p(Q, X) = \sum_{l=1}^{L} p(Q, X, F_0 = l) \tag{8}$$

This would mean running the decoder over all the L different systems and summing their output. If the pitch frequency is a dynamic variable, it could be hidden by marginalizing the distribution $p(x_n, F_{0n}|q_n)$ over F_{0n} to obtain the emission distribution $p(x_n|q_n)$ and performing decoding according to (4) [2]. Again in hybrid HMM/ANN system it is not clear how to marginalize continuous valued pitch frequency. However, for the case of discrete valued pitch frequency, it could be hidden to estimate $p(x_n|q_n)$ in the following way:

$$p(x_n|q_n) = \sum_{l=1}^{L} p(x_n, F_{0n} = l|q_n) \tag{9}$$

$$\approx \sum_{l=1}^{L} p(x_n|q_n, F_{0n} = l) \cdot P(F_{0n} = l) \tag{10}$$

and performing decoding according to (4). Equation (10) corresponds to (7), when the pitch frequency is hidden. In an earlier study, we investigated the effectiveness of pitch frequency as static information. We did not observe any improvement in the performance of the system [8]. Hence, in this paper we restrict ourselves to the case where pitch frequency is treated as dynamic information.

3. Experiments

3.1. Systems

We study 3 different hybrid HMM/ANN systems.
Baseline: System using standard acoustic features based on (4).
System 1: System with x_n and F_{0n} based on (6); F_{0n} is continuous valued.
System 2: System with F_{0n} independent of q_n based on (7); F_{0n} is discrete 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 the PhoneBook speech corpus for speaker-independent task-independent, small vocabulary (75 words) isolated word recognition [9]. For the connected word recognition task, we use the OGI Numbers speech corpus which contains free-format numbers spontaneously spoken by different speakers [10]. The definitions of the training, validation, and evaluation sets are similar to [11] and [12], for the PhoneBook corpus and the 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 Mel frequency cepstral coefficients (MFCCs), the first-order derivatives (delta) of the ten MFCCs and the c_0 (energy coefficient) are extracted for each time frame, resulting in a 21 dimensional 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 perceptual linear prediction (PLP) coefficients 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 pitch frequency is extracted using simple inverse filter tracking (SIFT) algorithm [13]. A 5-point median smoothing is performed on the pitch frequency contour. We evaluated our pitch estimation algorithm on the Keele Pitch Database [1] [14]. The results of this evaluation are given in Table 1. It shows that the pitch frequency estimation is reliable. In future, we would like to improve it further using other pitch frequency estimation approaches. In case of the systems where pitch frequency is continuous valued, the pitch frequencies are normalized by the highest pitch frequency which is 400Hz in our case (same for all utterances). The normalization is done in order to avoid saturation of the sigmoids [15].

Table 1: Evaluation of pitch estimation algorithm for 5 male and 5 female utterances. Gross error = $\frac{n_c}{n_v}$ where n_c is the total number of comparisons for which the difference between estimated pitch frequency and reference pitch frequency is higher or lower than 20% of reference pitch frequency and n_v is the total number of comparisons for which estimated pitch frequency and reference pitch frequency represent voiced speech. AMD - Absolute mean deviation.

Gender	Voiced in error (%)	Unvoiced in error (%)	High gross error (%)	Low gross error (%)	AMD (Hz)
Female	6.5	2.9	1.1	16.0	3.7
Male	22.3	1.5	3.7	5.1	2.0

3.3. Experimental Studies

The PhoneBook systems were trained with the 21 dimensional MFCC features. The OGI Numbers systems were trained with the 39 dimensional PLP features. The baseline systems were trained with the standard acoustic features. The number of parameters of system trained on PhoneBook database and OGI database are 139K and 538K, respectively. We have trained different baseline systems by varying the size of the ANN, all of them yielding performance similar to the one quoted in this paper.

In case of System 1, we trained a multilayer perceptron (MLP) by concatenating the standard acoustic feature vector with the pitch frequency at every frame i.e. the input layer contains additional inputs corresponding to the pitch frequency. In this case, we would be taking advantage of the MLPs ability to estimate higher order correlation between the components of the input feature, e.g. [4, page 75]. The number of parameters of system trained on PhoneBook database and OGI Numbers database are 144K and 465K, respectively.

The System 2 was implemented in the following manner.

1. The pitch frequency contour is estimated for all the training utterances.
2. The pitch frequencies are then vector quantized into three discrete regions, where one of the discrete regions models the unvoiced speech.
3. An MLP corresponding to each of the discrete regions is trained by finding the nearest discrete region corresponding to the value of the pitch frequency at that frame. The only exception is that the silence regions are observed by all the three MLPs. This is done because silence regions are nonspeech regions.

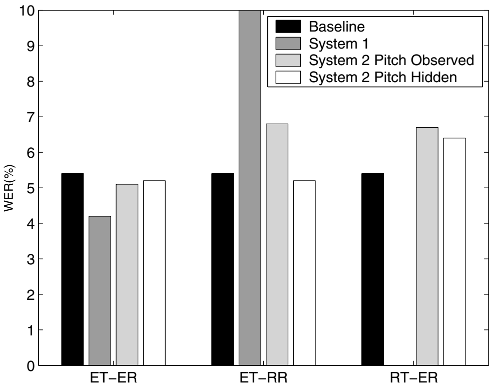

Figure 1: *Isolated word recognition on PhoneBook database. The performance is expressed in-terms of word error rate (WER). ET-ER: Systems trained with estimated F_{0n} and tested with estimated F_{0n}, ET-RR: Systems trained with estimated F_{0n} and tested with random F_{0n}. RT-ER: Systems trained with random F_{0n} and tested with estimated F_{0n} (results presently not available for System 1).*

During recognition, we study two strategies, namely, having the pitch frequency observed (O) and having the pitch frequency hidden (H). When the pitch frequency is observed during recognition, the single MLP corresponding to each observed F_{0n} is used. This is done on a frame-by-frame basis. When the pitch frequency is hidden, all the MLPs are used the according to (10). The number of parameters (sum of the parameters of all the 3 neural networks) of system trained on PhoneBook database and OGI Numbers database are 144K and 465K, respectively.

The results of the studies conducted on the PhoneBook database and OGI Numbers database are shown in Figure 1 and Figure 2, respectively (labelled ET-ER in Figures 1 and 2). In both the studies, System 1 and System 2 perform better than the baseline. In the case of PhoneBook system, the significant improvement (99% confidence) is observed for System 1, where as in case of OGI Numbers signification improvement (98% confidence) is observed for both System 1 and System 2.

In order to verify that the improvement in the performance of System 1 is due to pitch frequency and not due to increase in the input dimensionality or knowledge of voicing, we trained System 1 by concatenating the voicing decision with x_n i.e. substituting the pitch frequency value by 1 wherever pitch existed. The performances obtained were similar to the baseline. This suggests that improvement was not just due to the increase in the input dimension or voicing knowledge. We conducted two additional studies to investigate the role of pitch frequency during training and recognition. In the first study (labelled RT-ER in the Figures 1 and 2), we trained System 1 and System 2 with random pitch frequency values (within the range of the pitch frequency estimator) and conducted recognition with estimated pitch frequency values. In another study (labelled ET-RR in Figures 1 and 2), we conducted recognition experiments where the System 1 and System 2 were trained with estimated pitch frequency values and during recognition the estimated pitch frequency values were substituted by random pitch fre-

[1] ftp://ftp.cs.keele.ac.uk/pub/pitch/Speech

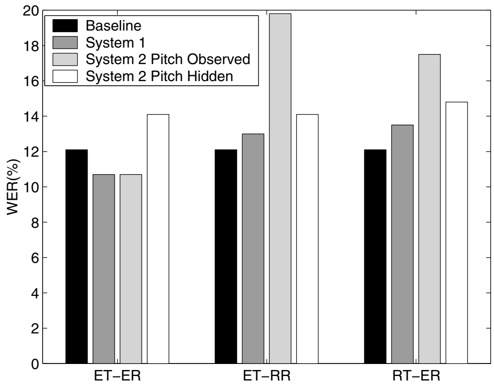

Figure 2: *Connected word recognition on OGI Numbers database. The performance is expressed in-terms of word error rate (WER). ET-ER: Systems trained with estimated F_{0n} and tested with estimated F_{0n}, ET-RR: Systems trained with estimated F_{0n} and tested with random F_{0n}. RT-ER: Systems trained with random F_{0n} and tested with estimated F_{0n}.*

quency values (within the range of the pitch frequency estimator). The results of this study are shown in Figures 1 and 2 for the PhoneBook Database and OGI Numbers Database, respectively. It can be observed from the results that the performance of the systems does not improve over the baseline system. This suggests that during training, the system has learned the relationship between the acoustic feature, the true estimate of the pitch frequency and the HMM states and this relationship is the one which is contributing towards the improved performance of the systems (labelled ET-ER in Figures 1 and 2) when the true estimate of the pitch frequency is observed.

4. Summary and Conclusion

In this paper, we studied two different ways in which pitch frequency can be incorporated in state-of-the-art hybrid HMM/ANN systems. Both approaches studied here performed better than the baseline system. System 1 yielded significant improvement for both the isolated word recognition task and connected word recognition task; whereas System 2 performed significantly better than the baseline for the connected word recognition task only. Our results suggest that pitch frequency can indeed help in improving the performance of ASR. The results obtained complements the recent efforts to model pitch frequency within the framework of HMM/GMM and dynamic Bayesian networks [1, 2].

In case of System 2, the difference between the performance of observed case and hidden case when random pitch frequencies were substituted for the estimated pitch frequencies (ET-RR case in Figures 1 and 2) shows the advantage of hiding the pitch frequency during recognition, when reliable estimate of pitch frequency is not available.

In the future, we would like to extend this study to incorporate other additional information such as rate-of-speech and short-time energy in the context of modelling speaker variability in spontaneous speech.

5. Acknowledgements

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

6. References

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

[2] T. A. Stephenson, J. Escofet, M. Magimai-Doss, and H. Bourlard, "Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables," in *NNSP*, 2002.

[3] H. Hermansky, "Perceptual linear predictive(PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1991.

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

[5] 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.

[6] S. L. Lauritzen and F. Jensen, "Stable local computations with conditional gaussian distributions," *Statistics and Computing*, vol. 11, no. 2, pp. 191–203, April 2001.

[7] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," in *Eurospeech*, 1993, pp. 1003–1006.

[8] Mathew Magimai.-Doss, Todd A. Stephenson, and Hervé Bourlard, "Modelling auxiliary information (pitch frequency) in hybrid HMM/ANN based ASR systems," IDIAP-RR 62, IDIAP, 2002.

[9] 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 *ICASSP*, 1995, pp. 1767–1770.

[10] R. A. Cole, M. Fanty, and T. Lander, "Telephone speech corpus at CSLU," in *ICLSP*, 1994.

[11] 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 *ICASSP*, 1767-1770, 1997, pp. 524–528.

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

[14] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Eurospeech*, 1995, pp. 837–840.

[15] Y. LeCun, L. Bottou, G. B. Orr, and K.-R Müller, "Effiecient BackProp," in *Neural Networks: Tricks of the Trade*, Genevieve N. Orr and Klaus-Robert Müller, Eds., chapter 1, pp. 9–50. Springer-Verlag, 1998.