

Optimization Tool for Speech Recognition

¹Kadam V.K

Department of Electronics, P.E.S College of Engineering , Nagsenvan, Aurangabad-431002 (M.S)

²R.C.Thool

Department of Information Technology, SGS Institute of Engg. Tech., Vishnupuri, Nanded 431606 (M.S)

¹email : vipulsangram@pescoe.ac.in, ²email : rcthool@ssgs.ac.in

Abstract

The proposal can be realized using HMM which is the universal model for the speech recognition. This proposal is an approach to increase the effectiveness of Hidden Markov Models (HMM) in the speech recognition field. This approach could determine the optimum topology with a practical computation time, and the performance can be comparable to the best recognition performance provided by the conventional maximum likelihood approach with manual tuning considering the decoding process. Thus, by using the proposed method, can automatically and rapidly determine an acoustic model topology with the highest performance, enabling us to dispense with manual tuning procedures when constructing acoustic models for speech recognition considering the decoding process in account. In this proposal I have compare the HMM with the Maximum Entropy Markov Model(MEMM), Factor analysed hidden Markov models (FAHMM) & Variational Bayesian Estimation and Clustering for Large Vocabulary Continuous Speech Recognition (VBEC) .All these models have tried to enhance the HMM for speech recognition. Here I have totally focus on the optimum area as shown in section 4 .This is optimized by using VBEC & HMM considering optimum topology for training data. In this proposal we have consider the effect of decoding condition on the optimality. Keeping this in view I have concentrated on optimum area.

I. Introduction

State-of-the-art speech and speaker recognizers typically extract cepstral features such as Mel Frequency Cepstral Coefficients (MFCCs) from consecutive frames of the speech signal, and use the MFCCs and their time derivatives as feature vectors in conjunction with a statistical classifier, such as a Hidden Markov Model (HMM) or a Gaussian Mixture Model (GMM). The MFCC features are calculated from the power spectrum, and include some harmonic structure related to the fundamental frequency, especially for high-pitched female speakers. For speaker recognition tasks, this property of MFCC features has been found to have a dramatic effect on accuracy. In speaker recognition evaluations conducted by the National Institute of Standards and Technology (NIST) in recent years it was found that speaker recognition systems perform considerably worse for high pitched speakers, and when the target speakers' pitch varies between enrollment and testing. Present speaker recognition systems exhibit extreme sensitivity to absolute pitch values and pitch mismatch. For example, speakers who exhibit a large pitch mismatch can experience an Equal Error Rate (EER) more than twice worse than pitch matched speakers. The effect of pitch-induced mismatch and variability in the features on speech recognition systems has received much less attention than pitch mismatch in speaker recognition. Variability due to pitch may be implicitly addressed by training a speech recognition system on a corpus collected

from a large, diverse collection of speakers. However, based on previous work using vocal tract length normalization and feature-space maximum-likelihood linear regression to reduce inter-speaker variability, we might expect that explicit reduction or elimination of pitch-induced feature variability could lead to better recognition performance.

Human beings are able to recognize speech amazingly well in high levels of background noise. On the other hand, the performance of automatic speech recognition (ASR) systems degrades dramatically with increasing noise. Part of the reason for this difference lies in the fact that the auditory system incorporates several features that make it more robust to noise. Most contemporary ASR systems attempt to incorporate some of these features. For instance, the most common feature representation of speech signals in ASR systems, the MEL spectrum, incorporates a simplified model of Critical band analysis, which resolves the speech signal into overlapping frequency bands of increasing width, with center frequencies spaced like the human auditory system. Another popular feature representation, the perceptual linear prediction or PLP spectrum, incorporates models of the equal loudness characteristic as well as the intensity-loudness relationship present in the auditory system when there is a mismatch between the acoustic conditions of training and application environments for a speech recognition system, the performance of the system very often is seriously degraded. Various sources give rise to this mismatch, such as additive noise, channel distortion, different speaker characteristics, different speaking modes, etc. The robustness of speech recognition techniques with respect to any of these different mismatched acoustic conditions thus becomes very important, and a variety of techniques have been developed to improve the system performance [3].

Speech recognition is a promising technique that may allow computers to discern human intentions. Recognition performance



Fig.1. State Transition in HMM: x - hidden states, y - observable outputs, a - transition probabilities, b - output probabilities plays an important role in this context, and depends strongly on the precise acoustic modeling of speech. Some of the most important applications of digital signal processing techniques have been in the area of speech processing. In fact, a large percentage of the theoretical 'background or digital signal

processing has been derived from speech studies and by speech researchers. As we shall see, digital processing has been applied to a wide range of problems in speech including spectrum analysis, channel vocoders, homomorphic processing systems, speech synthesizers, linear prediction systems; and computer voice response system .The basic assumption of almost all speech processing systems is that the source of excitation and the vocal tract system are independent. It is this source-system independence that allows us to discuss the transmission function of the vocal tract and to let it be excited by any of the possible sources to control the model above requires knowledge of the appropriate parameters (pitch period, switch position, amplitude, and filter coefficients) as a function of time. This is the goal of almost all speech analysis systems i.e., to estimate the appropriate model parameters from real speech. The goal of most speech synthesis systems is to use these parameters, obtained in any reasonable manner, to derive a synthetic speech signal that is indistinguishable perceptually from the original signal. Speech analysis-synthesis [1,6,7]. Applications of speech recognition may be for Automatic translation, Automotive speech recognition, Speech Biometric Recognition Dictation, Hands-free computing: voice command recognition computer user interface, Home automation etc.

II. Models Available For Speech Recognition

Statistical models for speech recognition usually model speech as a sequence of observations=O1O2.....O4 (acoustic features), produced by an unobservable "true" state sequence, S=S1S2...S4 (sequence of sub-phones, phonemes, or words). The states take on values in a finite state space V Within this framework, the goal of a speech recognizer is to find the state

sequence \hat{S} with the maximum posterior probability given an observed sequence O

$$\hat{S} = \underset{S}{\text{arg max}} \frac{p(S / O)}{p(S)} \tag{1}$$

We will discuss some few available following models for speech recognition considering optical modeling because recognition performance plays an important role in this context, and depends strongly on the precise acoustic modeling of speech.

A. Hidden Markov Models (HMMs)

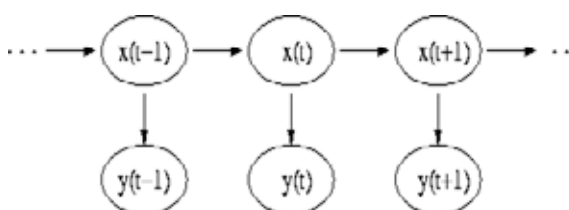


Fig.2 :General Architecture of an HMM.

Considering the acoustic modeling a category is expressed by a set of clustered-state triphone hidden Markov models (HMMs). In past decades, hidden Markov models (HMMs) have been studied extensively and have proved to be a very effective modeling technique. When solving (1) with the HMM, is obtained by maximizing the joint probability.

$$\hat{S} = \underset{S}{\text{arg max}} \frac{p(O/S) p(S)}{p(O)} = \underset{S}{\text{arg max}} p(O/S) p(S) \tag{2}$$

This approach uses Bayes' rule to compute $p(S/O)$ through a generative model $P(O/S)$. The HMM parameters defining $P(O/S)$ are typically estimated to maximize the likelihood of the training observations.

A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications. A HMM can be considered as the simplest dynamic Bayesian network. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, musical score and bioinformatics. Following fig.1 shows State Transition in HMM

III. Architecture of a hidden Markov model

The diagram above shows the general architecture of an HMM. Each oval shape represents a random variable that can adopt a number of values. The random variable $x(t)$ is the value of the hidden variable at time t. The random variable $y(t)$ is the value of the observed variable at time t. The arrows in the diagram (often called a trellis diagram) denote conditional dependencies.

From the diagram, it is clear that the value of the hidden variable $x(t)$ (at time t) only depends on the value of the hidden variable $x(t - 1)$ (at time t - 1). This is called the Markov property. Similarly, the value of the observed variable $y(t)$ only depends on the value of the hidden variable $x(t)$ (both at time t). The probability of observing a sequence of length L is given by

$$P(Y) = \sum_x P(Y / X)P(X) \tag{3}$$

Where the sum runs over all possible hidden node sequences although it is straight-forward to compute P(Y) by matrix multiplication of the transition matrix, doing so for many real-life problems requires potentially large amounts of compute storage and compute time. There exist a variety of algorithms that, while not providing exact results, do provide reasonably good results, with considerably less demand on storage and compute time. These include the forward-backward algorithm, and smaller still, the Viterbi algorithm. What constitutes "good enough" is often application-dependent: thus, for example, in radio communications, a well-designed application of the Viterbi algorithm will provide results within a fraction of a decibel of the exact results; a fraction of a decibel is "good enough" in this context [9].

B. Hidden Markov model (HMM)-based speech recognition

Modern general-purpose speech recognition systems are generally based on (HMMs). This is a statistical model which outputs a sequence of symbols or quantities. One possible reason why HMMs are used in speech recognition is that a speech signal could be viewed as a piece-wise stationary signal or a short-time stationary signal. That is, one could assume

in a short-time in the range of 10 milliseconds, speech could be approximated as a stationary process. Speech could thus be thought as a Markov model for many stochastic processes (known as states).

Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, to give the very simplest set up possible, the hidden Markov model would output a sequence of n-dimensional real-valued vectors with n around, say, 13, outputting one of these every 10 milliseconds. The vectors, again in the very simplest case, would consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short-time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have, in each state, a statistical distribution called a mixture of diagonal covariance Gaussians which will give likelihood for each observed vector. Each word, or each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes.

Described above are the core elements of the most common, HMM-based approach to speech recognition. Modern speech recognition systems use various combinations of a number of standard techniques in order to improve results over the basic approach described above. A typical large-vocabulary system would need context dependency for the phones it would use cepstral normalization to normalize for different speaker and recording conditions; for further speaker normalization it might use vocal tract length normalization (VTLN) for male-female normalization and maximum likelihood linear regression (MLLR) for more general speaker adaptation. The features would have so-called delta and delta-delta coefficients to capture speech dynamics and in addition might use heteroscedastic linear discriminant analysis (HLDA); or might skip the delta and delta-delta coefficients and use splicing and an LDA-based projection followed perhaps by heteroscedastic linear discriminant analysis or a global semitied covariance transform (also known as maximum likelihood linear transform, or MLLT). Many systems use so-called discriminative training techniques which dispense with a purely statistical approach to HMM parameter estimation and instead optimize some classification-related measure of the training data. Examples are maximum mutual information (MMI), minimum classification error (MCE) and minimum phone error (MPE).

Decoding of the speech would probably use the Viterbi algorithm to find the best path, and here there is a choice between dynamically creating a combination hidden Markov model which includes both the acoustic and language model information, or combining it statically beforehand (the finite state transducer, or FST, approach).

C. Factor analysed hidden Markov models

The factor analysed HMM is introduced in a generative model framework. Factor analysis is a statistical method for modelling the covariance structure of high dimensional data using a small number of latent (hidden) variables. It is often used to model the data instead of a Gaussian distribution with full covariance matrix. Factor analysis can be described by the following generative model.

$$x \sim N(O, I)$$

$$O = Cx + v, v \sim N(\mu^{(o)}, \Sigma^{(o)}) \quad (4)$$

Where x is a collection of k factors (k -dimensional state vector) and O is a p -dimensional observation vector. The covariance structure is captured by the factor loading matrix (observation matrix), C , which represents the linear transform relationship between the state vector and the observation vector. The mean of the observations is determined by the error (observation noise) modelled as a single Gaussian with mean vector an

$\mu^{(o)}$ and diagonal covariance matrix $\Sigma^{(o)}$. The observation process can be expressed as a conditional distribution,

$$p(O/x) = N(O, Cx + \mu^{(o)}, \Sigma^{(o)}).$$

Also, the observation distribution is a Gaussian with mean vector $\mu^{(o)}$ and covariance matrix

$$CC^T + \Sigma^{(o)}.$$

The baseline HMM system was produced by standard iterative mixture splitting using four iterations of embedded training per mixture configuration until no decrease in the word error rate was observed. The word error rates with the number of free parameters per HMM state up to 6 components are presented on the 1st row in Table 3, marked HMM. The best performance was 3.76% obtained with 10 mixture components. The number of free parameters per HMM state in the best baseline system

was $\eta = 780$ per state. However, increasing the number of mixture components in a standard HMM system can be seen to model the intra-frame correlation better. A FAHMM system with state space dimensionality $k=39$ and a global observation matrix, denoted as GFAHMM, was built for comparison. The global full 39 by 39 observation matrix was initialised to an identity matrix and the variance elements of the single mixture baseline HMM system were evenly distributed between the observation and state space variances. The number of state

space components was set to one, $M^{(x)}=1$ and the observation space components were increased by the mixture splitting procedure. The system corresponds to a global full loading matrix SFA with non-identity state space covariance matrices. The number of additional free parameters per state was 39 due to the state space covariance matrices, which could not be subsumed into the global observation matrix, and 1521 globally due to the observation matrix. Nine full iterations of embedded training were used, each with 20 within iterations. The results are presented on the third row in Table 3, marked GFAHMM. The best performance, 3.68%, was achieved with 5 mixture components. The deference in the number of free parameters

between the best baseline, $M^{(o)}=10$, and the best GFAHMM

system, $M^{(o)}=5$, was 351 per state. However, the GFAHMM system provides a relative word error rate reduction [10].

D. Maximum Entropy Markov Model (MEMM)

In contrast, direct modeling attempts to model the posterior Probability $P(S/O)$ directly. This approach has been used for statistical natural language understanding, for information extraction and segmentation, and only recently for acoustic modeling. There are many potential advantages as well as challenges for direct modeling. The direct model can potentially

make decoding simpler. It can also be a joint acoustic and language model. However, joint estimation would require a large amount of parallel speech and text data, clearly a challenge for data collection. The direct model allows for the potential combination of multiple sources of data in a unified fashion; for example, asynchronous and overlapping features can be incorporated formally, unlike the case for HMMs. Thus, it will be possible to take advantage of suprasegmental features like prosodic features and a multitude of other features such as acoustic phonetic features, speaker style, rate of speech, channel differences, etc. We expect features from different levels of linguistic hierarchies e.g., and to play an important role in direct modeling. Historically direct modeling was considered difficult due to the lack of a developed theoretical framework and foreseen computational difficulties. However, with recent advances in exponential models and increases in computational power, this approach has become practical. Exponential models are based on the maximum entropy principle and were successfully applied to classification problems. Recently, McCallum *et al.* modeled sequential processes using a direct model similar to the HMM in graphical structure and used exponential models for transition- observation probabilities. It was called a maximum entropy Markov model (MEMM) [2]-[3].

Although certain algorithms have been proposed for dealing with complicated model structure (model topology), they require heuristic tuning since they are based on the maximum likelihood (ML) criterion. That is to say, since the likelihood value increases monotonically as the number of model parameters increases, ML always leads to the selection of the model structure with the largest number of parameters, and this approach cannot determine the model topology appropriately without using heuristic tuning. Some partially successful approaches to the automatic determination of the acoustic model topology have been reported that use such information criteria as minimum description length or Bayesian information criterion (MDL1). The conventional MDL based approaches simplistically interpret the structure of the acoustic model as just a set of many single Gaussians, even though the structure is complicated and parameters could depend on each other in reality. Also the effects from the latent variables in HMMs and GMMs are excluded due to such an extreme simplification. Therefore, the description length cannot accurately represent the complexity inherent in acoustic model topology, and the MDL cannot determine the acoustic model topology properly [1].

E.Variational Bayesian Estimation and Clustering for Large Vocabulary Continuous Speech Recognition (VBEC)

In this the automatic determination of the optimum topology in an acoustic model by using Gaussian Mixture Model (GMM)-based phonetic decision tree clustering and an efficient model search algorithm utilizing the acoustic model characteristics. The proposal was realized using a Variational Bayesian Estimation and Clustering for speech recognition (VBEC) framework. This approach could determine the optimum topology with a practical computation time, and the performance was comparable to the best recognition performance provided by the conventional maximum likelihood approach with manual tuning.

The effectiveness of these methods has also been shown for much harder tasks, such as a spontaneous speech recognition task and a very large vocabulary speech recognition task

(vocabulary of 1.8 million) in. VBEC can automatically and rapidly determine an acoustic model topology with the highest performance, enabling us to dispense with manual tuning procedures when constructing acoustic models. In this the effectiveness of MSINGLE and MMIXTURE compared with the straightforward VBEC implementation, VBEC 2-phase search and the ML manual method. Here we compare these approaches not only in terms of the model topology but also in terms of computation time. To consider the experiments in more detail, in this they used isolated word recognition tasks to test various situations. The experimental conditions are summarized in Table 4. The training data consisted of about 3000 Japanese sentences (4.1 h) spoken by 30 males. They prepared two test sets as with the previous LVCSR experiments, each of which consisted of 100 Japanese city names spoken by 25 males (a total of 1200 words), as shown in Table 3. First, they have described the straightforward implementation of VBEC using VB iterative calculation within the original VBEC framework to prepare in-band models. In this experiment, they have approximated the VBEC iterative method by fixing the frame-to-state alignments during splitting and by using a phonetic decision tree construction as well as MSINGLE and MMIXTURE. Even in this situation, the full version of the iterative algorithm is unrealistic because of the VB iterative calculation in GMM. Namely, they used 45 personal computers with state-of-the-art specifications, so that the computation for each phonetic decision tree could be carried out in parallel (we call the VBEC iterative method within the original VBEC framework VBEC AMP (Acoustic Model Plant) because it is finally realized by such a large number of computers). Moreover, in order to reduce the computation time needed for the iterative calculation, they employed an approximation to reduce the

Table 1

Word error rates (%) and number of free parameters, η , on the RM task, versus number of mixture components for the observation pdfs, for HMM, and GFAHMM systems.

	$M^{(o)}$	1	2	3	4	5	6
HMM	η	78	156	234	312	390	468
	Wer[%]	7.79	6.68	5.05	4.32	4.09	3.99
GFAHMM	η	117	195	273	351	429	507
	Wer[%]	6.52	4.88	4.28	3.94	3.68	3.77

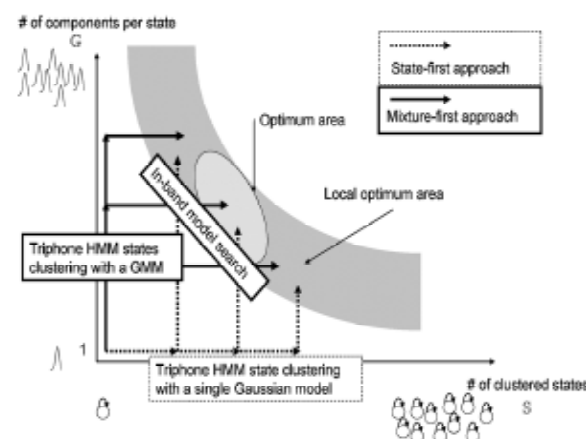


Fig. 3

decision branches when choosing the appropriate phonetic question. They have first derived the 10 best questions from 44 questions by applying all the questions to state splitting with a single Gaussian based state clustering method, which did not require any iterative calculations. Then, the iterative calculations were carried out for the derived 10 best questions. Here the trial suggested that the questions selected when using the 10 best questions covered about 95% of those selected when using all the questions, and were sufficient when carrying out iterative calculations for all the GMMs to construct a set of clustered-state triphone HMMs. Finally, the optimum model was also determined from the in-band models using the in-band model search as well as MSINGLE and MMIXTURE. As with the LVCSR experiments, we prepared a total of acoustic models for the ML manual method, and a total of 10 sets of clustered-state HMMs (1, 5, 10, 15, 20, 25, 30, 35, 40, and 50 components) for the VBEC automatic methods. The obtained model topology, performance and computation time for constructing acoustic models are listed in Table 4.

III. Proposed Work

At present, VBEC automatic determination only considers the optimum topology for training data, and does not consider the effects of the decoding condition on the optimality (e.g., mixture and state pruning process and the search parameter dependence in decoding) for the optimum model topology. We need further study on the superiority from a practical point of view. In future work, I would like to extend the model determination taking the decoding process into consideration, for example, by reflecting the effect of the decoding strategy of the mixture or state pruning process in the objective function of our optimized tool for speech recognition. We have further focus on how to optimize the optimum area as shown in following fig, considering the decoding process.

The VB objective function

$$F^m = \left\langle \log \frac{p(O, S, V / \Theta, m) p(\Theta / m)}{\bar{q}(\Theta / O, m)} \right\rangle_{\frac{\bar{q}(\Theta / O, m)}{\bar{q}(S, V / O, m)}} - \left\langle \log \bar{q}(S, V / O, m) \right\rangle_{\bar{q}(S, V / O, m)} \quad (5)$$

By separating F^m into two components: one is composed solely of $\bar{q}(S, V / O, m)$, whereas the other is mainly composed of $\bar{q}(\Theta / O, m)$. Therefore, we define F_{Θ}^m and $F_{S, V}^m$, and represent F^m as follows:

$$\begin{cases} F_{\Theta}^m \equiv \left\langle \log \frac{p(O, S, V / \Theta, m) p(\Theta / m)}{\bar{q}(\Theta / O, m)} \right\rangle_{\frac{\bar{q}(\Theta / O, m)}{\bar{q}(S, V / O, m)}} \\ F_{S, V}^m \equiv \left\langle \log \bar{q}(S, V / O, m) \right\rangle_{\bar{q}(S, V / O, m)} \end{cases} \quad (6)$$

Table III

Experimental conditions for isolated word recognition

Sampling rate/Quantization	16 kHz/16 bit
Feature vector(24 dimensions)	12-order MFCC with Δ MFCC
Window	Hamming
Frame size/shift	25/10 ms
Number of states	3 (Left to right)

Number of phoneme categories	27
Number of phonetic questions	44

Table IV

Comparison with Iterative & Non-Iterative State Clustering. #S: number of clustered states, #G: number of components per Gaussian

	ML-manual {#S,#G}	2-phase search			
{#S,#G}	AMP				
{#S,#G}	MSINGLE				
{#S,#G}	MMIXTURE				
{#S,#G}					
Model topology	{500,30}, {500,35}	{2,642,5}	{548, 30}	{253, 35}	{224, 35}
Recognition rate (%)	97.1,97.6	96.3,94.9	97.3, 98.0	97.6, 98.2	97.8, 97.8
Time(hour)	244	56	1,150	30	37

where $\bar{q}(S, V / O, m)$ is Posterior distribution model parameters, $\bar{q}(\Theta / O, m)$ is Output distribution parameters,

$p(O, S, V / \Theta, m)$ is Output distribution represent a phoneme

acoustic model, $p(\Theta / m)$ is Model parameters, O is data set, S is set of sequence of HMM states, V is set of sequence of Gaussian Mixture Component.

Using above object function it is possible determine appropriate model topology. In determination of appropriate model we deal with the determination of the HMM-contextual topology, whose search spaces are conFig.d by the search spaces of the HMM state clustering, and GMM topology and . The other aspects, the HMM-temporal and feature vector topologies, are regarded as invariants. That is to say, we focus on the two dominant topological aspects that have much wider search spaces than the other two, and that, by changing, largely affect the recognition performance. Then we consider the determination

of model topology \bar{m} in a subspace of the original search space defined is as follows:

$$\bar{m} = \arg \max_{m \in (S \times G)} F^m \quad (7)$$

Since we employ appropriate clustering algorithms, the HMMcontextual and GMM topologies can be represented by one-dimensionalized indexes, the numbers of clustered HMM states and GMM components per state, as shown in Fig. 3. In order to realize the optimum model topology, we utilize the two characteristics of the acoustic model, the inverse-proportion band and the unimodality. By preparing a number of acoustic models in the band, and by choosing the model that has the best VBEC objective function score, we can determine the optimum model topology, as shown in Fig. 3 (in-band model search).

References

[1] Shinji Watanabe, Atsushi Sako, and Atsushi Nakamura, "Automatic Determination of Acoustic Model Topology Using Variational Bayesian Estimation and Clustering for Large Vocabulary Continuous Speech Recognition", IEEE

- Trans. Audio, Speech, and Language Processing, vol. 14, no. 3, May 2006, pp.855-871.
- [2] Hong-Kwang Jeff Kuo, and Yuqing Gao, "Maximum Entropy Direct Models for Speech Recognition" IEEE Trans. Audio, Speech, and Language Processing, vol. 14, no. 3, May 2006, pp. 873.
- [3] Ran D. Zilca, Brian Kingsbury, Navrátil, & Ganesh N. Ramaswamy, "Pseudo Pitch Synchronous Analysis of Speech With Applications to Speaker Recognition", IEEE Trans. Audio, Speech, and Language Processing, vol. 14, no. 2 March 2006, pp.467-468.
- [4] Jethran Guinness, Bhiksha Raj, Bent Schmidt-Nielsen, Lorenzo Turicchia, Rahul Sarpeshkar "A Companding Front End for Noise-Robust Automatic Speech Recognition", ICASSP 2005 pp1-2.
- [5] Jieh-Wei Hung, Member, IEEE, and Lin-Shan Lee, Fellow, IEEE "Optimization of Temporal Filters for Constructing Robust Features in Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, no. 3, May 2006
- [6] Lawrence R. Rabiner, Bernard Gold "Theory & Application of Digital Signal Processing", PHI Eleventh Edition Jan. 1998, pp 658-667.
- [7] P.Ramesh Babu, "Digital Signal Processing", SPIPT fourth Edition July 2007, pp 10.1-10.11.
- [8] McCallum & Freitag and Pereira, "Maximum entropy Markov Model" Preliminary CRF work Published in 2000.
- [9] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, 77 (2), February 1989 pp 257-286.
- [10] A-V.I. Rosti & M.J.F. Gales, "Factor analysed hidden Markov models for speech recognition", Proceedings of Computer Speech and Language 16 October 2003 pp 4-20.



Kadam V.K
Associate Professor & Research Student,
Department of Electronics, Head of Knowledge Center & Cisco
Networking Academy
P.E.S College of Engineering
Nagsenvan, Aurangabad-431002 (M.S)
email:vipulsangram@pescoe.ac.in
Dr.Babasaheb Ambedkar Marathwada University, Aurangabad-
431002 (MS)



R.C.Thool
Professor & Head
Department of Information Technology
SGGS Institute of Engineering Technology
Vishnupuri, Nanded 431606 (M.S)
email:rcthool@sngs.ac.in
(An autonomous institute set up and 100% funded by
Government of Maharashtra)