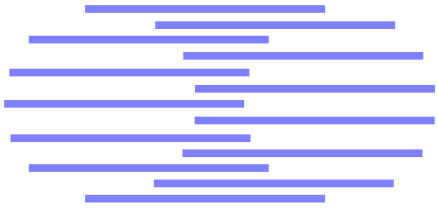
IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse



Pronunciation models and their evaluation using confidence measures

Mathew Magimai Doss [a]
Hervé Bourlard [a,b]

IDIAP–RR 01-29

October 2001

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, Martigny, Switzerland
[b] EPFL, Lausanne, Switzerland

IDIAP Research Report 01-29

Pronunciation models and their evaluation using confidence measures

Mathew Magimai Doss Hervé Bourlard

October 2001

Abstract. In this report, we present preliminary experiments towards automatic inference and evaluation of pronunciation models based on multiple utterances of each lexicon word and their given baseline pronunciation model (baseform phonetic transcription). In the present system, the pronunciation models are extracted by decoding each of the training utterances through a series of hidden Markov models (HMM), first initialized to only allow the generation of the baseline transcription but iteratively relaxed to converge to a truly ergodic HMM. Each of the generated pronunciation models are then evaluated based on their confidence measure and their Levenshtein distance with the baseform model. The goal of this study is twofold. First, we show that this approach is appropriate to generate robust pronunciation variants. Second, we intend to use this approach to optimize these pronunciation models, by modifying/extending the acoustic features, to increase their confidence scores. In other words, while classical pronunciation modeling approaches usually attempt to make the models more and more complex to capture the pronunciation variability, we intend to fix the pronunciation models and optimize the acoustic parameters to maximize their matching and discriminant properties.

1 Introduction

The goal of Automatic Speech Recognition (ASR) is to produce a word sequence M that best matches an acoustic vector sequence $X = \{x_1, x_2, \ldots, x_N\}$. In case of statistical ASR, this problem is then formulated in terms of finding the model M^* such that:

$$M^* = \underset{M}{\operatorname{argmax}}\ P(M|X) \approx \underset{M}{\operatorname{argmax}}\ P(X|W, \Theta_A)P(M|\Theta_L), \tag{1}$$

where $P(X|M, \Theta_A)$ is the *acoustic model* (AM) and $P(M|\Theta_L)$ is the *language model* (LM) (a priori probability of the utterance M). The acoustic model parameters Θ_A are usually estimated using a training corpus of acoustic data with corresponding phonetic transcriptions (generally obtained from a pronunciation dictionary), and the language model parameters Θ_L are usually estimated from a large amount of text corpora.

While the performance of the resulting system strongly depends on the amount of (AM and LM) training data, its also depends very much on the adequacy of the model, including, e.g., the topological (lexical) constraints of the models and the assumptions about the probability density functions. Indeed, *maximum likelihood* (ML) training is known to be optimal only if the models are correct.

In the case of the acoustic model $P(X|M, \Theta_A)$, the adequacy of the model is often improved by relaxing the assumptions regarding the emission probability density functions (e.g., going from Gaussian to multi-Gaussian distributions, or to artificial neural networks) and by using context-dependent phonetic units. However, this is often not enough to capture all the variability of the speech signal and it is also necessary to relax the lexical constraints (allowed phonetic transcriptions associated with each lexicon word). Indeed, phonological studies of the way a word is pronounced in different lexical contexts often lead to more than one acceptable pronunciation for many words. Furthermore, casual human conversation exhibits an even larger amount of nonstandard variability in pronunciation, in which case the phonetic realization of a word can depend on many parameters such as speaking rate, pitch, etc.

To improve the quality of the acoustic models, different pronunciations (baseforms) of each word are usually compiled into pronunciation dictionaries and most speech recognizers use such a dictionary for pronunciation modelling. To introduce multiple pronunciations into the acoustic model, equation (1) may be modified as[1]:

$$\begin{aligned} M^* &= \underset{M}{\operatorname{argmax}}\ P(X|M)P(M) \\ &= \underset{M}{\operatorname{argmax}} \sum_{B \in \mathcal{B}(M)} P(X, B|M)P(M) \\ &\approx \underset{M}{\operatorname{argmax}} \sum_{B \in \mathcal{B}(M)} P(X|B)P(B|M)P(M) \end{aligned} \tag{2}$$

where $\mathcal{B}(M)$ is the set of possible baseforms associated with the word sequence M, B is a specific baseform (pronunciation), $P(B|M)$ is the probability to associate baseform B with word string M.

During recognition, we will then search for the model M maximizing (2) by summing over all possible baseforms of M. Replacing the sum operator in (2) by a max operator, a usual practice is also to simply match the acoustic signal against all possible pronunciations of all the words in a candidate word string and pick the word string (and thus baseform sequence) yielding the maximum likelihood.

As addressed in the present report, an important research issue is to further reduce the mismatch between the dictionary representation of words and their actual realization by using an improved pronunciation model. In current ASR systems, this is often achieved simply by adding many pronunciation alternatives for each word. These variants are often defined on the basis of a priori phonological rules

[1] And dropping the parameters Θ_A and Θ_L for the sake of clarity.

and/or on maximum likelihood training, adding pronunciation variants such that the likelihood of some acoustic training data is increased [SKW98]. However, while this improves the matching properties of each of the lexicon individually, the way these multiple pronunciations are defined is also known to increase the confusability between the words.

Instead of making the acoustic models more complex as briefly described above (with the risk of increasing confusability), the present report was motivated by the goal of using additional auxiliary (acoustic) information to improve the matching properties of a single lexical baseform (e.g., standard phonetic transcription). It was also motivated by the fact that standard baseform inference algorithms usually do not make use of the a priori knowledge of the standard phonetic transcription. In this report, we however mainly focused on measuring the adequacy of a given baseform model by (1) relaxing the lexical constraints of the baseline phonetic transcription and (2) measuring the confidence level of the acoustic match for the inferred baseforms:

1. Inference of pronunciation variants: as usually done, this is achieved by phonetically decoding each training utterance through an ergodic HMM. However, in our case, this ergodic HMM is initialized to only allow the generation of a first order approximation of the baseline phonetic transcription, and is later relaxed to iteratively converge towards a fully ergodic HMM. For each of these HMM configurations, a phonetic transcription is generated and evaluated. If the inferred phonetic transcription does not diverge too quickly from the standard baseform when relaxing the HMM topology, this means that the pronunciation model and the acoustic observation match well.

2. Evaluation of each of the inferred phonetic transcriptions through the use a confidence measure and the Levenshtein distance between the inferred phonetic sequence and the associated baseline transcription. In this report, we thus also discuss the different confidence measures studied to evaluate the adequacy of the baseform pronunciation models.

All the work reported here has been done in the context of hybrid HMM/ANN ASR, using artificial neural networks (ANN) to estimate local posterior probabilities used as HMM emission probabilities [BM94].

In the following section, we briefly describe the inference process based on the relaxation of the baseline phonetic transcription constraints. In Section 3, we present the different types of HMM/ANN-based confidence measures and discuss/evaluate their potential use in pronunciation modeling. Finally, Section 4 presents the initial experimental studies that were carried out to evaluate the baseform pronunciation models. In the future, we intend to complement the acoustic features with additional auxiliary information [SMB01] to improve the adequacy of the baseline pronunciation model, which will be measured in terms of the evaluation criteria discussed here (confidence measure and Levenshtein distance). This is further discussed in the conclusion section.

2 Baseform Pronunciation Inference

HMM inference is a technique to infer the best HMM model associated with a given utterance. This inference is performed by doing phonetic/subword-level decoding of the utterance, matching the acoustic sequence X on an *ergodic* HMM model M_g. For our studies we use hybrid HMM/ANN system [BM94] and each HMM state q_k correspond to a phone which is associated with a particular ANN output. A perfect ergodic HMM model contains a set of fully-connected phonetic states (with a minimum duration constraint for each phoneme) with uniform transition probabilities. Figure 1 shows a 3-state ergodic HMM model, including the non-emitting initial and final states I and F.

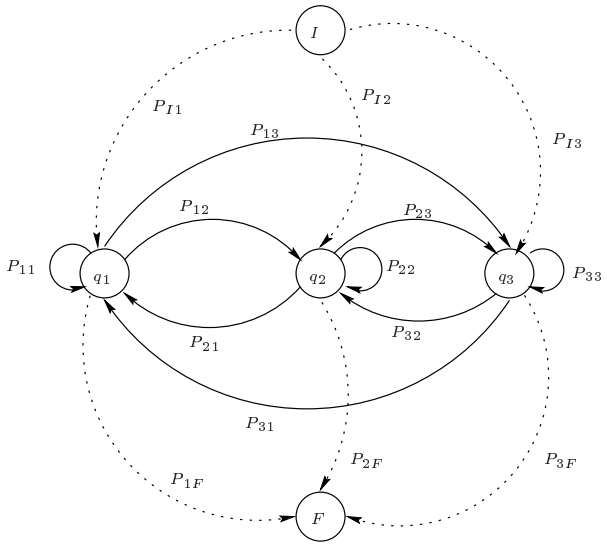


Figure 1: 3-state Ergodic HMM

The transition probability matrix for this ergodic HMM is

$$T = \begin{bmatrix} P_{II} & P_{I1} & P_{I2} & P_{I3} & P_{IF} \\ P_{1I} & P_{11} & P_{12} & P_{13} & P_{1F} \\ P_{2I} & P_{21} & P_{22} & P_{23} & P_{2F} \\ P_{3I} & P_{31} & P_{32} & P_{33} & P_{3F} \\ P_{FI} & P_{F1} & P_{F2} & P_{F3} & P_{FF} \end{bmatrix} \quad (3)$$

$$= \begin{bmatrix} 0.00 & 0.33 & 0.33 & 0.33 & 0.00 \\ 0.00 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.00 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.00 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix} \quad (4)$$

A perfect ergodic HMM is capable of producing any state sequence (since there is no grammar or lexical constraint in it), as opposed to left-to-right HMM which can only produce constrained state sequences.

In ASR system, a specific lexicon word is often represented in terms of a phonetic sequence, referred to as the *baseline phonetic transcription*. In the case of the above 3-phone set, a word could be represented, e.g., by its phonetic transcription $\{q_2, q_1, q_2\}$, thus represented as a left-to-right HMM, as shown in Figure 2.

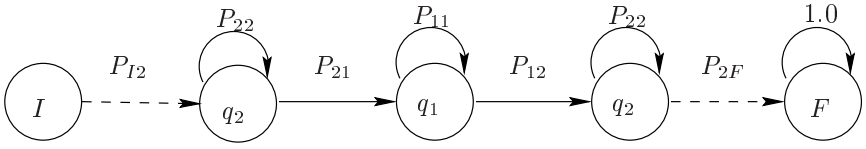


Figure 2: Left-to-Right HMM

In standard ASR systems, the baseline phonetic transcription is often complemented by pronunciation variants that are obtained by performing phonetic decoding (inference) of several utterances of the same word through the ergodic HMM. This process thus generates multiple phonetic transcriptions

which are merged into a single (more complex) HMM or kept separate. These pronunciation variants can also be pruned/smoothed to keep only the most representative ones, and thus avoid confusion with other lexicon words. In this inference process though the knowledge of the baseline transcription is not explicitly used and it is often the case that the inferred baseforms diverge a lot from the baseline.

To alleviate this problem, we decided to study here a new way to infer phonetic variants (baseforms). For each lexicon word, and given its baseline phonetic transcription, we first start from a transition matrix representing a first-order approximation of the baseline transcription (thus only allowing the transitions present in the left-to-right HMM). This transition matrix is referred here to as a *constrained ergodic model*. This *constrained transition matrix* is then slowly relaxed by adding an ϵ value to each of its entries, followed by a re-normalization, and thus yielding a transition matrix such as in (5). For very small ϵ values, this model is nearly equivalent (in a first-order approximation) to the baseline form, while for large values of ϵ the model is then equivalent to a truly ergodic HMM. In the case of Figure 1, the relaxed transition probability matrix would take the form given in (5).

$$T = \begin{bmatrix} 0.0 & \frac{\epsilon}{1+3\epsilon} & \frac{1+\epsilon}{1+3\epsilon} & \frac{\epsilon}{1+3\epsilon} & 0.0 \\ 0.0 & \frac{\epsilon}{1+4\epsilon} & \frac{1+\epsilon}{1+4\epsilon} & \frac{\epsilon}{1+4\epsilon} & \frac{\epsilon}{1+4\epsilon} \\ 0.0 & \frac{1+\epsilon}{2+4\epsilon} & \frac{\epsilon}{2+4\epsilon} & \frac{\epsilon}{2+4\epsilon} & \frac{1+\epsilon}{2+4\epsilon} \\ 0.0 & \frac{\epsilon}{4\epsilon} & \frac{\epsilon}{4\epsilon} & \frac{\epsilon}{4\epsilon} & \frac{\epsilon}{4\epsilon} \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \tag{5}$$

We note here that when a constrained ergodic HMM is used for inference, it can still recognize state sequences other than the baseline pronunciation. For example, the above example of constrained ergodic HMM can recognize state sequences such as $\{q_2, q_1, q_2, q_1, q_2\}$ or just q_2 apart from the intended state sequence $\{q_2, q_1, q_2\}$. This is because of the first order Markov model assumption and also an ergodic HMM does not model the temporal information like the left-to-right HMM.

A constrained ergodic HMM encodes the lexical constraint information through the transitional probability matrix. But when the ϵ value is increased the lexical constraint is relaxed such that the transition probability matrix starts allowing transitions which are not present in the baseline pronunciation. The perfect ergodic HMM is a special case of relaxed ergodic HMM which does not have any lexical constraint information.

The underlying idea exploited in the present report thus consists in generating for each utterance of a given lexicon word several phonetic transcriptions through successive relaxation of the transition matrix. The quality of these inferred phonetic transcriptions will then be assessed in terms of different measures, including:

1. Confidence measure: measuring the confidence level (based on posterior probabilities or likelihood ratio) between the acoustic sequence and the different inferred baseforms, as already proposed in [WR98].

2. Levenshtein score between the inferred and the baseline transcription.

3. "Speed of divergence": if the inferred phonetic transcription does not diverge too quickly from the baseline transcription when relaxing the transition probability matrix, it is indeed a sign that the baseline transcription is quite "stable", and thus reliable.

In the following section, we describe the different confidence measures used for our studies.

3 Confidence Measures

Hybrid HMM/ANN based systems are capable of estimating the posterior probability $P(M|X)$ of model M given the acoustic observations, X [BM94]. It is also known that posteriors by themselves are measure of confidence. In literature, different confidence measures that can be derived from a

hybrid HMM/ANN system based on local phone posterior probabilities, $P(q_k|x_n)$ have been suggested [Wil99], where x_n is the feature vector at time frame n and q_k is the state hypothesis.

We are going to investigate two confidence measures, namely, scaled-likelihood-ratio and posterior probability based confidence measure. These confidence measures are briefly described below, where b_k and e_k are the begin and end frames of a state hypothesis q_k:

3.1 Scaled Likelihood-Ratio Confidence Measure

Likelihood ratio is a very common hypothesis testing technique [SL96, Fre71]. Likelihood ratio of a sequence X hypothesized as model M is:

$$L(X|M) = \frac{p(X|M)}{p(X|\bar{M})} \tag{6}$$

where $\bar{M}$ represents the hypothesis that X is not associated with M. It is difficult to estimate $p(X|\bar{M})$ directly, so it is often estimated by $p(X|M_g)$, where M_g is a model that can generate all possible state sequences, *e.g.* a perfect ergodic HMM.

$$L(X|M) \simeq \frac{p(X|M)}{p(X|M_g)} \tag{7}$$

The log-likelihood ratio is defined as

$$\mathcal{L}(X|M) \simeq \log p(X|M) - \log p(X|M_g) \tag{8}$$

Since we are working with hybrid HMM/ANN systems, we only have access to scaled-likelihoods obtained by dividing the local posterior probabilities by the class prior probabilities.

$$\frac{p(x_n|q_k)}{p(x_n)} = \frac{P(q_k|x_n)}{P(q_k)} \tag{9}$$

Where the class prior probabilities $P(q_k)$ are estimated from the relative frequencies of the phone labels in the acoustic training data. The scaled-likelihood of a segment assigned to state q_k in the hypothesis is defined as:

$$SL(q_k) = \prod_{n=b_k}^{n=e_k} \frac{P(q_k|x_n)}{P(q_k)} \tag{10}$$

Minus log-scaled-likelihood of the segment can be expressed as:

$$\mathcal{SL}(q_k) = -\sum_{n=b_k}^{n=e_k} \log(\frac{P(q_k|x_n)}{P(q_k)}) \tag{11}$$

The probability of a decoding hypothesis is always underestimated due to the observation independence assumption. This underestimate creates a bias towards shorter decoding hypotheses. Duration normalization counteracts this bias. The duration normalized log-scaled-likelihood is given as:

$$\mathcal{SLN}(q_k) = \frac{\mathcal{SL}(q_k)}{e_k - b_k + 1} \tag{12}$$

The *word-level log-scaled-likelihood* is defined by averaging $\mathcal{SLN}(q_k)$ over the constituent phones as given below.

$$\mathcal{SLN}_w(X|M) = \frac{\sum_{k=1}^{k=K} \mathcal{SLN}(q_k)}{K} \tag{13}$$

where K is the number of phones in hypothesis M. The *word-level log-scaled-likelihood ratio* is then defined as:

$$\mathcal{SLR}_w \simeq \mathcal{SLN}_w(X|M) - \mathcal{SLN}_w(X|M_g) \tag{14}$$

Lower this value, higher the confidence level is.

3.2 Posterior Probability based Confidence Measure

The posterior based confidence measure is defined as the normalized logarithm of the segment-based accumulated posterior probabilities.

For a given segmentation (resuting in our case from a Viterbi algorithm using local posterior probabilites), we define the accumulated posteriors for all the acoustic vectors observed on state q_k as:

$$CM_{post}(q_k) = \prod_{n=b_k}^{n=e_k} P(q_k|x_n) \tag{15}$$

Defining minus log of $CM_{post}(q_k)$ as the state-based confidence measure

$$\mathcal{CM}_{post}(q_k) = -\sum_{n=b_k}^{n=e_k} \log P(q_k|x_n) \tag{16}$$

the *normalized word-level posterior probability based confidence measure* is then defined as:

$$\mathcal{CM}_{wpost} = \frac{1}{K}\sum_{k=1}^{k=K} \frac{\mathcal{CM}_{post}(q_k)}{e_k - b_k + 1} \tag{17}$$

Lower this value, higher the confidence level is.

3.3 Confidence based on Levenshtein score

When HMM inference is performed, we obtain the phonetic decoding from the best path. The confidence measures are computed using the best path as described earlier in this section. In our case apart from confidence measures, measuring the difference between the phonetic decoding and the baseform pronunciation is of equal interest because it is possible to get a phonetic decoding different from the baseform pronunciation. We express the difference between the phonetic decoding obtained from HMM inference and the baseform pronunciation in terms Levenshtein distance. Given two strings, the Levenshtein distance is defined as the minimum number of changes that has to be made in one string to convert it into another string. Consider two strings /c/ /a/ /t/ and /a/ /c/ /t/, in this case the Levenshtein score is two as a minimum of two changes have to be made to convert any one of the string into another. We call the Levenshtein distance as Levenshtein score in this report. The core idea behind this study is that if the acoustic observations and the model match well then

1. When the inference is performed on a constrained ergodic HMM the confidence is high and Levenshtein score is zero.
2. As the constrained ergodic HMM is relaxed to perfect ergodic HMM the confidence and the Levenshtein score does not change.

Confidence measures can be used to evaluate pronunciation models and identify poor pronunciation models. In [WR98], confidence measure $\mathcal{CM}_{post}(q_k)$ was used to evaluate the baseform and alternate pronunciations of each word in the lexicon. The frequency of occurrence of the pronunciations and the confidence measure together were used to create a new lexicon. This approach improved the performance of the system. In the later sections, we describe our approach to evaluate baseform pronunciation model using confidence measures.

4 Preliminary Experimental Results

4.1 Database and Experimental Setup

For our studies, we used a speaker-independent MLP trained on a subset of *PolyPhone* database [CCC+96]. The MLP contained 234 input units (representing nine consecutive acoustic frames with a feature vec-

tor of size 26 each), 600 hidden units and 36 output units (associated with 36 phones). The feature vector consisted of 13 plp and 13 delta-plp features. The phonetic recognition rate of the MLP on a *PolyPhone* test set was above 85% [ACB97].

Seventeen isolated words spoken by 13 different speakers (on average) was chosen from *PolyVar* database [CCC+96] for the baseform evaluation studies. The phone level performance of the MLP on this subset of words was 77.6%. The baseform pronunciation of all the words starts with a silence phone and ends with a silence phone. For example, the exact baseform pronunciation of the word *annulation* is
[*aa*] [*nn*] [*uu*] [*ll*] [*aa*] [*ss*] [*yy*] [*on*]
But in our case the baseform pronunciation is represented as
[*sil*] [*aa*] [*nn*] [*uu*] [*ll*] [*aa*] [*ss*] [*yy*] [*on*] [*sil*].
This is basically done to accommodate the silence in the beginning and at the end of the utterance.

4.2 Experimental Studies

In this section, we present our experimental studies. Sections 4.2.1 and 4.2.2 describe the baseform pronunciation evaluation using $\mathcal{CM}_{wpost}$ and $\mathcal{SLR}_w$, respectively. Section 4.2.3 presents the experiments carried out to study the effect of duration modeling on the confidence measures.

4.2.1 Evaluation of baseforms and relaxed baseforms using posterior-based confidence measure

In this study, we used $\mathcal{CM}_{wpost}$ as given in (17) as the confidence measure to evaluate pronunciation models. The state duration for each phone was five.

Our approach for generating and evaluating pronunciation models was the following:

1. For a given word utterance X, and given its known baseline phonetic transcription, initialize the $K \times K$ transition probability matrix (with $K = 38$ in our case, corresponding to the 36 phones, plus initial and final states) with a very small ϵ value, to constraint the ergodic model to be equivalent to a first-order approximation of the baseline phonetic transcription of the word.

2. Perform forced Viterbi decoding based on that model using local posterior probabilities $P(q_k|x_n)$.

3. From the resulting best path, extract the phonetic level decoding and compute $\mathcal{CM}_{wpost}$.

4. Compute Levenshtein score between the phonetic sequence obtained from step 3 and the actual baseline pronunciation.

5. Relax the underlying model towards a fully ergodic model by increasing the ϵ value, and repeat steps 2-4 to infer new phonetic baseforms and compute their associated confidence measure.

Table 1 illustrates on 3 different utterances (numbered 4, 5 and 10) of the same word ("*concert*") the evolution of the confidence measure and the Levenshtein distance when iteratively relaxing the transition constraints in the underlying Markov model. From this table, we can observe that the phonetic decoding of utterance 10 diverges from the baseform pronunciation faster than utterances 4 and 5. This seems to indicate that the acoustic observations of utterance 10 matches poorly with the acoustic model as compared to that of utterances 4 and 5. This fact is also reflected in the confidence scores obtained for the constrained ergodic HMM (e.g., for $\epsilon = 10^{-20}$ and comparing the first row of columns 2, 4 and 6).

The results obtained for different utterances of the word *concert* are given in Table 2, where $\mathcal{CM}^r_{wpost}$ and LS^r are the confidence score and Levenshtein score obtained for a relaxed ergodic HMM. The confidence score is expressed in *negative logarithmic* scale. So a low score means that the confidence is high and a high score means that the confidence is low.

Table 1: Normalized posterior-based confidence measures $\mathcal{CM}_{wpost}$ and Levenshtein scores LS obtained for different values of ϵ for utterances 4, 5 and 10 of word *concert*. Small $\mathcal{CM}$ values reflect high confidence levels. Good match between baseline pronunciation model and acoustic utterance is characterized by a high confidence level and stability (low LS) when increasing ϵ.

ϵ	$\mathcal{CM}^4_{wpost}$	LS^4	$\mathcal{CM}^5_{wpost}$	LS^5	$\mathcal{CM}^{10}_{wpost}$	LS^{10}
10^{-20}	0.624061	0	0.241951	0	0.723885	0
10^{-16}	0.624061	0	0.241951	0	0.723885	0
10^{-10}	0.624061	0	0.241951	0	0.636117	1
10^{-5}	0.365516	4	0.241951	0	0.479075	2
10^{-3}	0.344603	4	0.241951	0	0.404442	3
10^{-1}	0.314058	4	0.241951	0	0.404442	3
10^{0}	0.314058	4	0.196920	1	0.404442	3
10^{1}	0.314058	4	0.196920	1	0.404442	3
10^{2}	0.314058	4	0.196920	1	0.404442	3

The results in Table 2 show that as the constrained ergodic HMM is relaxed, the confidence score generally decreases (compare columns 1 and 3) but at the same time the decoding hypothesis also changes this is indicated by the change in the Levenshtein score (compare columns 2 and 4). There are utterances for which the confidence score and the decoding hypothesis have remained same (row 5). It means that the acoustic observations of those utterances and acoustic model have matched well, whereas, for other utterances there has been mismatch between the acoustic observation and the model. The degree of mismatch is indicated by the Levenshtein score. In our case, the confidence score and the Levenshtein score together are thus good measure of how well the acoustic observations match the acoustic model. For most of the utterances, the inferred phonetic transcription diverges from the baseline transcription as the constrained ergodic HMM was relaxed to fully ergodic HMM (increasing the value of ϵ), as reflected through an increase of the Levenshtein score. However, from Table 1, it is clear that this divergence rate is not the same across different utterances.

4.2.2 Evaluation of baseforms and relaxed baseforms using scaled likelihood confidence measure

In this section, we describe baseform pronunciation model evaluation using log-scaled-likelihood ratio $\mathcal{SLR}_w$ as given in (14). The approach for evaluating the baseform pronunciation is the following.

1. For a given word and the baseline transcription, Initialize the $K \times K$ transition probability matrix with a small ϵ value.

2. Compute scaled-likelihoods for each frame.

3. Perform Viterbi decoding on the ergodic HMM. Get the phonetic decoding of the utterance and the confidence score for the best path.

4. Compute the Levenshtein distance between the phonetic decoding obtained in step 3 and the baseform pronunciation.

5. Re-initialize the transition probability matrix with a large ϵ (say 1000) and repeat steps 2 and 3.

6. Subtract the score obtained for the best path in step 5 from the score obtained in step 3. This score is the $\mathcal{SLR}_w$.

Table 2: Comparison of the posterior-based confidence measures $\mathcal{CM}_{wpost}$ and Levenshtein score (LS) for the word *concert* in the case of the baseline phonetic transcription (left columns) and the relaxed HMM (using $\epsilon = 10^{-1}$).

$\mathcal{CM}_{wpost}$	LS	$\mathcal{CM}^r_{wpost}$	LS^r
0.502744	0	0.312323	3
0.452068	0	0.334207	3
0.525777	0	0.352027	2
0.624061	0	0.314058	4
0.241951	0	0.241951	0
0.342735	0	0.219196	1
0.697103	0	0.349461	3
0.257648	0	0.200218	2
0.443140	0	0.403126	2
0.723885	0	0.404442	3
0.483746	0	0.352943	1
0.429921	0	0.215865	3
0.570243	0	0.287076	3

7. Increase the ϵ value, re-initialize the transition probability matrix and perform steps 2-6.

A duration of 5-states was chosen for each phone. The results obtained for the same utterances of word *concert* are shown in the Table 3. A score of 0.0000 means that the decoding hypothesis of the perfect ergodic HMM and the relaxed or constrained ergodic HMM are same. In practise, the likelihood of the perfect ergodic HMM is more than the relaxed or constrained ergodic HMM. But when the word level confidence score is computed, sometime it was observed that the confidence for the phonetic sequence obtained for a relaxed or constrained ergodic HMM is more than that of the phonetic sequence obtained for a perfect ergodic HMM. This means that the phonetic decoding obtained in case of such relaxed or constrained ergodic HMM is better than that of the phonetic decoding obtained for the perfect ergodic HMM. The negative scores in the Table 3 reflects this effect.

By comparing the results obtained in Study I and Study II, the following conclusions can be drawn

a. In general, the two confidence measures agree with each other.

b. The confidence measure and the Levenshtein score decrease or remain same when the model is relaxed.

c. The confidence score and Levenshtein score together are good measure of how good the acoustic observations match the acoustic model.

4.2.3 Duration constraint modeling

In ASR, minimum phone duration constraints are often used as additional knowledge to improve the overall recognition performance [Wan97] Generally, the phone duration is fixed before training or recognition. In the following, we studied the effect of different phone duration constraints $(2, 3, 5)$ on the confidence measures of the constrained and relaxed HMMs. In these studies, we used both the confidence measures $\mathcal{CM}_{wpost}$ as given in (17) and $\mathcal{SLR}_w$ as given in (14). Table 4 shows the results obtained by performing HMM inference on a relaxed ergodic HMM ($\epsilon = 10^{-1}$) for the different utterances of the word *concert*. The superscript denotes the phone duration.

Where ever the confidence scores for different durations are not matching it means that the decoding hypothesis has changed. From the results, it is clear that the confidence score and the Levenshtein score

Table 3: Comparison of the log-scaled-likelihood based confidence measures $\mathcal{SLR}$ and Levenshtein score (LS) for the word *concert* in the case of the baseline phonetic transcription (left columns) and the relaxed HMM (using $\epsilon = 10^{-1}$).

$\mathcal{SLR}_w$	LS	$\mathcal{SLR}_w^r$	LS^r
0.221790	0	0.106987	1
0.563138	0	-0.051750	3
0.108723	0	0.000000	2
0.751288	0	0.000000	4
-0.079178	0	-0.079178	0
0.096404	0	0.000000	1
0.809505	0	0.000000	1
0.326141	0	0.000000	4
0.264877	0	0.000000	2
0.734458	0	0.000000	5
0.517407	0	0.000000	5
0.492740	0	-0.063747	3
0.785493	0	0.000000	3

obtained depend upon the phone duration. In the literature, different approaches have been suggested to model phone duration through transition probabilities. In [Sie95], two sets of state-transition models were used. The state-transition probabilities in the first set were made equal to ignore any information present. The second set of models had transition probabilities adapted to the fast speech utterances. When tested on fast speech, it was observed that for models using adapted transition probabilities the performance improved and for the models using equal transition probabilities the performance decreased. In our studies, we explicitly modelled the phone duration through states. By varying the phone duration we are mainly varying the transition probability.

5 Summary, Discussion and Future Work

In this report, we introduced new ways to infer pronunciation variants and to evaluate their relevance. For each lexicon word to be modeled, the general idea is to start from a "constrained" ergodic model, corresponding to the first-order approximation of the baseline phonetic transcription and thus only allowing the generation of phonetic sequences basically identical to the baseline transcription. This constrained model is then iteratively relaxed to converge towards an actual ergodic HMM, thus allowing all possible phonetic sequences, within the usual minimum phone duration constraints. For each configuration of this relaxed ergodic HMM, the optimal phonetic sequence is extracted and its relevance is estimated in terms of (1) a confidence measure (based on posterior probabilities or scaled-likelihood) and (2) the Levenshtein distance with respect to the baseline transcription. A good pronunication model should result in a high confidence measure and a good stability when relaxing the ergodic HMM.

The preliminary studies reported here show that the acoustic confidence measures, $\mathcal{CM}_{wpost}$ and $\mathcal{SLR}_w$ together with Levenshtein score are good measure of how good the acoustic observations match the acoustic model. It is also clear from the studies that if the acoustic observations and the acoustic model match well then the confidence should be high and the Levenshtein score should be zero when HMM inference is performed on the constrained ergodic HMM and they should not change when the ergodic HMM is relaxed. The results obtained show that the acoustic observations and the acoustic do not match well all the time. This is primarily due to the variations introduced by the speaker/s

Table 4: Results obtained for different utterance of the word *concert* for different state durations 5, 3, and 2 subscript denotes the confidence measure and the superscript denotes the state duration,

$\mathcal{CM}^5_{wpost}$	LS^5	$\mathcal{CM}^3_{wpost}$	LS^3	$\mathcal{CM}^2_{wpost}$	LS^2
0.312323	3	0.336848	2	0.306587	2
0.334207	3	0.420334	0	0.352794	1
0.352027	2	0.168693	3	0.168693	3
0.314058	4	0.314058	4	0.314058	4
0.241951	0	0.230518	0	0.230518	0
0.219196	1	0.219196	1	0.219196	1
0.349461	3	0.349461	3	0.349461	3
0.200218	2	0.185385	2	0.185385	2
0.403126	2	0.343156	2	0.343156	2
0.404442	3	0.404442	3	0.404442	3
0.352943	1	0.326830	1	0.326830	1
0.215865	3	0.215865	3	0.221177	2
0.287076	3	0.287076	3	0.287076	3

such as different speaking rate, physiological differences, stress, different dialects *etc* (in the context of pronunciation modelling) [Haz98].

In literature, different approaches have been suggested to adapt the models so as to reduce the effect of these variation on the performance of the system [Haz00]. In previous research, recognition accuracy has been improved by adapting the acoustic model specific to auxiliary information such as gender, pitch and environment [Ace93, Liu94]. For example, gender modeling for automatic speech recognition uses the gender information to adapt the acoustic model [Liu94, NR97, Haz98]. During training a separate acoustic model ($p(X|M, A)$, where A is the auxiliary information *i.e.* gender) is trained for male and female speakers and during recognition the gender information is either used or inferred automatically. This approach has lead to improvement in the performance of the system [NR97]. Recently, similar studies have been reported in literature to model other discrete or continuous auxiliary information to improve the performance of the speech recognition system. Todd *et. al.* used pitch frequency as the auxiliary information with in the framework of Bayesian networks to improve the performance of ASR system [SMB01]. In this study, the pitch frequency was a discrete variable. Fujinaga *et. al.* [FNSS01] and Tuerk *et. al.* [TY01] have proposed approaches to model continuous auxiliary information with in the framework of HMM system. The results of Study III suggests that the transition probabilities does contain useful information for ASR and by conditioning them on auxiliary information the performance of the system may be improved. Thus, our future goal is to model auxiliary information such as speaking rate and pitch frequency to reduce the mismatch between the acoustic observations and the acoustic model, in other words to get better pronunciation models.

References

[ACB97] J. M. Anderson, G. Caloz, and H. Bourlard. Swisscom AVIS project (no. 392) advanced vocal interfaces services technical report for 1997. Technical Report RR-97-06, IDIAP, Martigny, Switzerland, December 1997.

[Ace93] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers, Boston MA, 1993.

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

[CCC+96] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, and Ph. Langlais. Swiss french polyphone and polyvar: Telephone speech databases to model inter- and intra- speaker variability. Technical Report rr-96-01, IDIAP, Martigny, Switzerland, April 1996.

[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.

[Fre71] John E. Freund. *Mathematical Statistics*. Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971.

[Haz98] Timothy J. Hazen. *The Use of Speaker Correlation Information for Automatic Speech Recognition*. PhD dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, January 1998.

[Haz00] Timothy J. Hazen. A comparison of novel techniques for rapid speaker adaptation. *Speech Comm.*, 31:15–33, 2000.

[Liu94] F.-H. Liu. *Environmental Adaptation for Robust Speech Recognition*. PhD dissertation, Carnegie Mellon University, Department of Electrical and Computer Engineering, December 1994.

[NR97] C. Neti and S. Roukos. Phone-context specific gender dependent acoustic-models for continuous speech recognition. In *ASRU*, December 1997.

[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.

[SKW98] Helmer Strik, Judith M. Kessens, and Mirjam Wester. *Proceedings of Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*. Automatic Acoustic Recognition Technologies, Dept. of Speech and Language, University of Nijmegen, The Netherlands, 1998.

[SL96] R. A. Sukkar and C-H. Lee. Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. *IEEE Trans. Speech, Audio Processing*, 4:420–429, 1996.

[SMB01] Todd A. Stephenson, M. Mathew, and H. Bourlard. Modeling auxiliary information in Bayesian network based ASR. In *Proceedings of Eurospeech*, pages 2765–2768, Aalborg, Denmark, September 2001.

[TY01] A. Tuerk and S. J. Young. Indicator variable dependent output probability modelling via continous posterior functions. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 473–476, 2001.

[Wan97] Xue Wang. *Incorporation Knowledge on Segmental Duration in HMM-Based Continuous Speech Recognition*. PhD dissertation, University of Amsterdam, Faculteit Der Letteren, April 1997.

[Wil99] David Arthur Gethin Williams. *Knowing What You Don't Know: Roles for Confidence Measures in Automatic Speech Recognition*. PhD dissertation, University of Sheffield, Department of Computer Science, Sheffield, May 1999.

[WR98] G. Williams and S. Renals. Confidence measures for evaluating pronunciation models. In *Proc. of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 151–155, Rolduc, Netherlands, May 1998.