IDIAP RESEARCH REPORT

On the Adequacy of Baseform Pronunciations and Pronunciation Variants

Mathew Magimai.-Doss [a] [b]
Hervé Bourlard [a] [b]

IDIAP–RR 04-27

May 2004

SUBMITTED FOR PUBLICATION

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] Dalle Molle Institute for Perceptual Artificial Intelligence, CH-1920 Martigny, Switzerland
[b] Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

IDIAP Research Report 04-27

On the Adequacy of Baseform Pronunciations and Pronunciation Variants

Mathew Magimai.-Doss Hervé Bourlard

May 2004

submitted for publication

Abstract. This paper presents an approach to automatically extract and evaluate the "stability" of pronunciation variants (i.e., adequacy of the model to accommodate this variability), based on multiple pronunciations of each lexicon words and the knowledge of a reference baseform pronunciation. Most approaches toward modelling pronunciation variability in speech recognition are based on the inference (through an ergodic HMM model) of a pronunciation graph (including all pronunciation variants), usually followed by a smoothing (e.g., Bayesian) of the resulting graph. Compared to these approaches, the approach presented here differs by (1) the way the models are inferred and (2) the way the smoothing (i.e., keeping the best ones) is done. In our case, indeed, inference of the pronunciation variants is obtained by slowly "relaxing" a (usually left-to-right) baseform model towards a fully ergodic model. In this case, the more stable the model is, the less the inferred model will diverge from it. Hence, for each pronunciation model so generated, we evaluate their adequacy by calculating the Levenshtein distance of the the new model with respect to the baseform, as well as their confidence measure (based on some posterior estimation), and models with the lowest Levenshtein distance and highest confidence are preserved. On a large telephone speech database (Phonebook), we show the relationship between this "stability" measure and recognition performance, and we finally show that automatically adding a few pronunciation variants to the less stable words is enough to significantly improve recognition rates.

1 Introduction

In standard automatic speech recognition (ASR) systems during recognition, for each acoustic observation x_n at time frame n, the acoustic model outputs the likelihoods of each subword unit e.g. phoneme, which is used by the subsequent decoding step. The decoding step uses the acoustic probabilities /likelihoods, the pronunciation model of the words present in the lexicon, and the language model (grammar) to output the most probable sequence of words that could have generated the acoustic observation sequence $X = \{x_1, \cdots, x_n, \cdots, x_N\}$ [BM94].

The lexicon of an ASR system contains the words and their standard pronunciations i.e. a sequence of subword units [Ost99]. We refer to this sequence of subword units of a word as baseform pronunciation of the word. The baseform pronunciation of each word is generally obtained from a standard lexical dictionary which contains both the meaning of the word and the way the word is to be pronounced. This could be further enriched by phonological rules. In standard hidden Markov model (HMM) based ASR during decoding, it is a stochastic pattern matching problem where given the acoustic models of subword units, we have to match the acoustic observation sequence X and pronunciation model (sequence of subword units). In speech recognition systems, it is generally expected that speaker(s) pronounce the words according to the phonetic transcription given in the lexicon; but speaker(s) do introduce pronunciation variation which leads to a mismatch between the acoustic observation and pronunciation model. The pronunciation variation can occur at the [SC99]

1. Acoustic characteristic level due to speaking style, speaking rate, different accent, pitch, differences in the length of the vocal tract, background noise (Lombard effect), emotion or stress.

2. Lexical characteristic level due to phonological processes such as assimilation, co-articulation, reduction, deletion and insertion, accent or "*liaisons*" in French.

For the reasons described above, the baseform pronunciation cannot properly model the pronunciation variation. Sometimes even with high frame/phoneme level performance the word performance can still be poor because the lexical constraints are not correct.

There are different ways to improve the match between acoustic parameters and pronunciation models, such as,

1. Adapting or enriching the pronunciation models. For example, generating new pronunciation variants and adding them to the lexicon or creating pronunciation lattices [SC99].

2. Adapting the acoustic model, such as iterative training [SC99], sharing the parameters of the phoneme models in baseform pronunciation with parameters of the phonemes in alternate realization(s) [Sar00].

3. Extracting subword units and word pronunciations automatically from the data [BO99].

The most common practice is to generate new pronunciation variants. The approaches used for generating new pronunciation variants can be broadly classified as, (a) knowledge-based (b) data-driven approaches, or (c) a mix of both [SC99]. The generated pronunciations are kept separate [SC99] or merged into a single (more complex) HMM [SO94]. These pronunciation variants can also be pruned/smoothed to keep only the most representative ones. However, while this improves the matching properties of each of the words individually, the way these multiple pronunciations are defined is also known to increase the confusion between words.

In this paper, we take an alternate approach where:

- The adequacy of the baseform pronunciation of words in the lexicon is evaluated [MDB01] i.e. how stable the baseform pronunciation is to acoustic variability.

- When the baseform pronunciation is inadequate for a given word, pronunciation variants that are as stable as possible to acoustic variability and at the same time not too dissimilar to the baseform pronunciation are extracted and added to the lexicon.

The adequacy of a given baseform pronunciation model is evaluated by (1) relaxing the lexical constraints of the baseform pronunciation and (2) measuring the confidence level of the acoustic match for the inferred pronunciation variants:

1. Inference of pronunciation variants: as is 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 baseform pronunciation, and is later relaxed iteratively to converge towards a fully ergodic HMM. For each of these HMM configurations, a phonetic transcription is generated (pronunciation variant) and evaluated.

2. Evaluation of each of the inferred phonetic transcriptions through the use of a confidence measure and the Levenshtein distance between the inferred phonetic sequence and the associated baseform pronunciation. Here, we basically assess the "stability" of the baseform pronunciation to perturbations through the confidence measure and Levenshtein distance obtained.

As a by product, this evaluation procedure provides a framework to extract new pronunciation variants which are reliable and closer to the baseform pronunciation. Investigation of the proposed approach on a task-independent speaker-independent isolated word recognition task has yielded significant improvement in the performance of the system.

In addition to this, we show that the proposed evaluation procedure can be used to evaluate different acoustic models. For example in this study, we use acoustic models that are trained with standard Mel frequency cepstral coefficients (MFCCs) acoustic features and acoustic models that are trained with both MFCCs and auxiliary features [MDSB03, SMDB04]. We observe that modelling standard acoustic features along with auxiliary features such as pitch frequency and short-term energy improves the stability of the baseform pronunciation of words.

The paper is organized as follows. Section 2 describes the pronunciation model evaluation procedure. Section 3 describes briefly the measures used in our studies to assess the adequacy of pronunciation. Section 4 and Section 5 present the experimental setup and analytical studies, respectively. Section 6 presents the pronunciation variants extracting procedure and the results of the recognition studies. Finally, Section 7 concludes with discussion and future directions of work.

2 Evaluation of Pronunciation Models

HMM inference is a technique to infer the "best" HMM model associated with a given set of utterances [MJ98]. This inference is done by performing subword-unit level decoding of the utterance, matching the acoustic sequence X on an *ergodic* HMM model. For our studies, we use hybrid hidden Markov model/artificial neural network (HMM/ANN) systems [BM94] and each HMM state q_k corresponds to a context-independent phoneme which is associated with a particular ANN output. A fully ergodic HMM model contains a set of fully-connected phonetic states with uniform transition probabilities. Figure 1 shows a 3-state ergodic HMM model, including the non-emitting initial and final states I and F.

A fully ergodic HMM is capable of producing any state sequence (since there is no grammar or lexical constraint in it), as opposed to a left-to-right HMM which can only produce constrained state sequences. An ergodic HMM is obviously too general to model lexical constraints. In current ASR systems, the words are usually represented as left-to-right sequences of subword-units. For example, Figure 2 illustrates a word represented by pronunciation $\{q_2, q_1, q_2\}$.

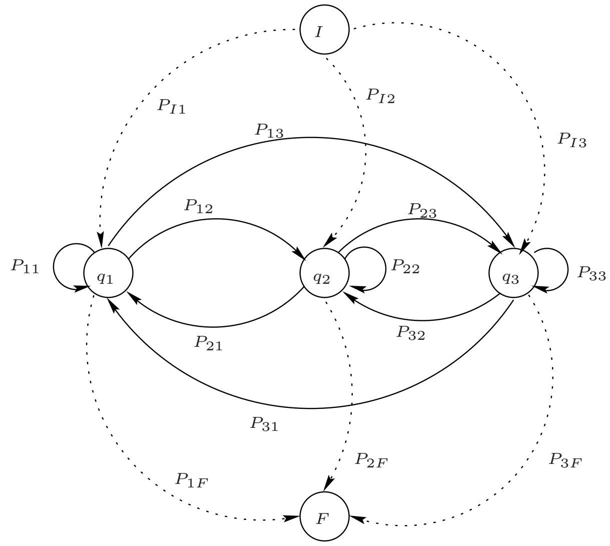

Figure 1: 3-state Ergodic HMM

The transition probability matrix for the fully ergodic HMM is

$$
\begin{aligned}
T &= \begin{bmatrix} P_{II} & P_{I1} & P_{I2} & P_{I3} & P_{IF} \\ P_{1I} & P_{11} & P_{12} & P_{13} & P_{1F} \\ P_{2I} & P_{21} & P_{22} & P_{23} & P_{2F} \\ P_{3I} & P_{31} & P_{32} & P_{33} & P_{3F} \\ P_{FI} & P_{F1} & P_{F2} & P_{F3} & P_{FF} \end{bmatrix} \\
&= \begin{bmatrix} 0.00 & 0.33 & 0.33 & 0.33 & 0.00 \\ 0.00 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.00 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.00 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix} \qquad (1)
\end{aligned}
$$

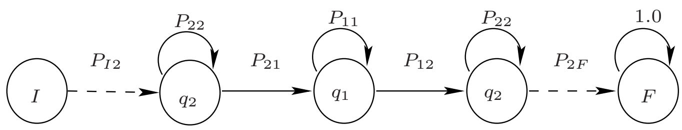

Figure 2: Left-to-Right HMM

In pronunciation modelling literature, the HMM inference approach is used to generate pronunciation variants [MJ98], by performing phonetic decoding (inference) of several utterances of the same/different words through the fully ergodic HMM which doesnot encode any lexical constraints. The proposed pronunciation model evaluation approach is based on a HMM inference mechanism which uses the prior knowledge of baseform pronunciation. For each lexicon word, and given its baseform pronunciation,

1. We first start from a transition matrix representing a first-order approximation of the baseform pronunciation (thus only allowing the transitions present in the left-to-right HMM). This is done by taking the transition probability of a fully ergodic HMM, say (1), adding an ϵ to the transitions present in the baseform pronunciation (for e.g. Figure 2) followed by a re-normalization, and

thus yielding a transition matrix such as in (2). This ergodic model is referred to here as a *constrained ergodic model*.

$$T = \begin{bmatrix} 0.0 & \frac{1}{3+3\epsilon} & \frac{1+3\epsilon}{3+3\epsilon} & \frac{1}{3+3\epsilon} & 0.0 \\ 0.0 & \frac{1}{4+4\epsilon} & \frac{1+4\epsilon}{4+4\epsilon} & \frac{1}{4+4\epsilon} & \frac{1}{4+4\epsilon} \\ 0.0 & \frac{1+4\epsilon}{4+8\epsilon} & \frac{1}{4+8\epsilon} & \frac{1}{4+8\epsilon} & \frac{1+4\epsilon}{4+8\epsilon} \\ 0.0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \tag{2}$$

For a large value of ϵ, this *constrained ergodic model* is a first order approximation of baseform pronunciation

2. This *constrained ergodic model* is then slowly relaxed by decreasing the value of ϵ. For $\epsilon = 0.0$, this model is then equivalent to a fully ergodic HMM.

We note here that when a constrained ergodic HMM is used for inference, it can still recognize state sequences other than the baseform pronunciation because of first order Markov assumption. For example, the above example of a 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 $\{q_2, q_1, q_2\}$.

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

The underlying idea exploited in the present paper thus consists of generating for each utterance of a given lexicon word several pronunciation variants through successive relaxation of the transition matrix, i.e. decreasing the value of ϵ. The quality of these inferred pronunciation variants are then 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 pronunciation variants, as already proposed in [WR99].

2. Levenshtein distance [SK99] between the inferred pronunciation variants and the baseform pronunciation.

3. "Speed of divergence": if the inferred pronunciation variants do not diverge too quickly from the baseform pronunciation when relaxing the transition probability matrix, it is indeed a sign that the baseform pronunciation is quite "stable", and thus adequate.

3 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]. 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 [WR99, MDB01], where x_n is the feature vector at time frame n and q_k is the state hypothesis.

In this paper, we use the posterior probability based confidence measure. The posterior based confidence measure is defined as the normalized logarithm of the segment-based accumulated posterior probabilities.

3.1 Confidence Measure

For a given segmentation (resulting 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:

$$CM_{post}(q_k) = \prod_{n=b_k}^{n=e_k} P(q_k|x_n), \tag{3}$$

where b_k and e_k are the begin and end frames of a state hypothesis q_k. Defining minus log of $CM_{post}(q_k)$ as the state-based confidence measure

$$\mathcal{CM}_{post}(q_k) = -\sum_{n=b_k}^{n=e_k} \log P(q_k|x_n) \tag{4}$$

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

$$\mathcal{CM}_{wpost} = \frac{1}{K}\sum_{k=1}^{k=K} \frac{\mathcal{CM}_{post}(q_k)}{e_k - b_k + 1}, \tag{5}$$

Where K is the number of constituent phonemes in the inferred model. *The lower the value $\mathcal{CM}_{wpost}$, the higher the confidence level is.*

3.2 Levenshtein Distance

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 inferred pronunciations and the baseform pronunciation is of equal interest because it is possible to infer a pronunciation with high confidence level that is completely different from the baseform pronunciation. We measure the difference between the inferred pronunciation and the baseform pronunciation in terms of Levenshtein distance (LD).

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 [SK99]. 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 strings into another.

4 Experimental Setup

We use the PhoneBook speech corpus for our studies [PFW+95]. There are 42 context-independent phonemes including silence, each modelled by a single emitting state. The standard acoustic vector x_n is the MFCCs extracted from the speech signal using an analysis 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. The auxiliary features used in this study are pitch frequency and short-term energy.

We use the following trained systems for our studies:

1. Hybrid HMM/ANN baseline system trained with standard features (system-base).

2. Hybrid HMM/ANN systems trained with standard features and auxiliary features. These systems have been shown to improve the performance of ASR systems [MDSB03, SMDB04]. The auxiliary features are used in two different ways

(a) Concatenated to the standard feature to get an augmented feature vector with which hybrid HMM/ANN system is trained. The system trained with pitch frequency as auxiliary feature is denoted as system-app-p, and the system trained with short-term energy as auxiliary feature is denoted system-app-e.

(b) Auxiliary features conditioning the emission distribution similar to gender modelling. The system trained with pitch frequency as the auxiliary feature is denoted as system-cond-p, and the system trained with short-term energy as the auxiliary feature is denoted system-cond-e.

For further details about how these systems can be implemented, refer to [MDSB03, SMDB04].

These systems were trained with a training set consisting of 19420 utterances and a validation set consisting of 7290 utterances. All the systems have the same number of parameters. The test set consists of 8 different sets of 75 word lexicon amounting 6598 utterances. These systems were trained for speaker-independent task-independent, small vocabulary (75 words) isolated word recognition. The words and speakers present in the training set, validation set and test set do not overlap.

5 Analytical Studies

We performed analytical studies using the acoustic models of system-base, system-app-p and system-cond-e. In our studies, system-app-p and system-cond-e perform significantly better than system-base. We used a part of the validation set, 75 words, each spoken on average by 12 different speakers. We performed evaluation of baseform pronunciations in the following manner:

1. For a given word utterance X, and given its known baseform pronunciation, initialize the $K \times K$ transition probability matrix (where $K = 44$ in our case, corresponding to the 42 phonemes, plus initial and final states) with a very large ϵ value (say 10^{10}), to constrain the ergodic model to be equivalent to a first-order approximation of the baseform pronunciation 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 distance LD between the phonetic sequence obtained from step 3 and the baseform pronunciation.

5. Relax the underlying model towards a fully ergodic model by decreasing the ϵ value, and repeat steps 2-4 to infer new phonetic transcription and compute their associated $\mathcal{CM}_{wpost}$ and LD.

The ideal case suitable for automatic speech recognition would be something like shown in Figure 3, where

1. When the inference is performed on a constrained ergodic HMM the Levenshtein distance is zero.

2. As the constrained ergodic HMM is relaxed to fully ergodic HMM the inferred pronunciations diverge less from the baseform pronunciation.

But in practice, we observe the following (see Figure 4)

1. When the inference is performed on a constrained ergodic HMM (i.e. large value for ϵ) the confidence level is low and the Levenshtein distance is low.

2. As the constrained ergodic HMM is relaxed to fully ergodic HMM the confidence level increases and the Levenshtein distance also increases.

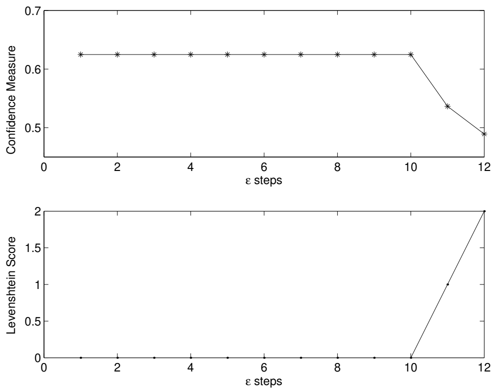

Figure 3: A case where the baseform pronunciation of word keeble uttered by a female speaker matches well with the acoustic observation. The inference was done with acoustic models of system-app-p.

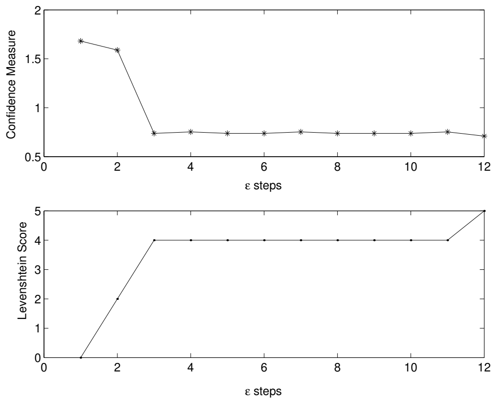

Figure 4: A case where the baseform pronunciation of word keeble uttered by a female speaker doesnot match well with the acoustic observation. The inference was done with acoustic models of system-app-p.

As can observed from the Figures 3 and 4 with posterior-based confidence and Levenshtein distance together we can analyze the stability of the baseform pronunciation; but it is difficult to make conclusions about the adequacy of the baseform pronunciation based on the evaluation of a single utterance. So, to make a decision about the adequacy of a baseform pronunciation and to extract pronunciation variants we will need more than one utterance.

We evaluated the baseform pronunciation of all the 75 words using their multiple utterances with the procedure described earlier in this section. We did the same evaluation with acoustic models of system-base, system-app-p and system-cond-e. The main outcomes of this analytical study are the following:

1. When the baseform pronunciation of a word matches acoustic observations well, the evaluation across different speakers mostly yields a behavior similar to Figure 3 i.e confidence level is high and when the lexical constraints are relaxed the speed of divergence is slow.

2. When a baseform pronunciation is inadequate, the confidence level is low and the speed of divergence is fast for most of the utterances of the word.

3. When comparing across the acoustic models system-base, system-app-p and system-cond-e, none of the models are totally superior over others. Most of the time the acoustic models trained with

MFCCs and auxiliary features match the baseform pronunciation well. In order to visualize it, at each inference step (i.e. for each value of ϵ) we combined the posterior-based confidence measure and the levenshtein score in the following way

$$comb = \mathcal{CM}_{wpost} + \log(1 + LD)$$

Taking $\log(1 + LD)$ is appropriate as LD is an integer and has a wide dynamic range compared to $\mathcal{CM}_{wpost}$. Also, we are interested in changes in LD at lower levels (i.e. the deviations that are not too far from the baseform pronunciation) which the log function represents well. A high *comb* value means low confidence i.e $\mathcal{CM}_{wpost}$ and/or LD are high. We observed that for the majority of the utterances it is low for the acoustic models trained with both MFCCs and auxiliary features when compared to acoustic models just trained on MFCCs. This is illustrated in Figure 5.

This is an interesting outcome i.e., the baseform evaluation procedure can be used to evaluate different acoustic models by fixing the pronunciation models. Also, an alternate approach to model pronunciation variation would be to fix the baseform pronunciations and optimizing acoustic parameters so as to maximize their matching and discriminating properties.

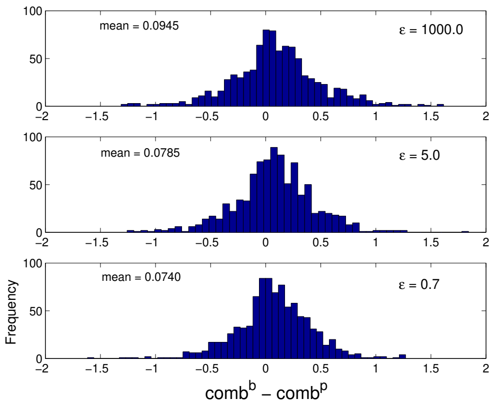

Figure 5: Histogram of difference between the $comb^b$ value (obtained by using acoustic models of system-base) and $comb^p$ value (obtained by using acoustic models of system-app-p) for different values of ϵ, for all the utterances.

6 Extraction of Pronunciation Variants and Speech Recognition Studies

We perform recognition studies by:

1. First evaluating the baseform pronunciations of the test lexicon words (these words are neither present in the training set or validation set) using the pronunciation evaluation procedure described in the last section.

2. For the words for which the baseform pronunciation are inadequate, pronunciation variants are extracted from the evaluation procedure itself and added to the lexicon.

3. Then, recognition studies are performed with the updated lexicon.

In order to do this study, we split the test set randomly (keeping the gender balance) into two parts: (a) "*H-set*", used for baseform pronunciation evaluation and pronunciation variant extraction (45% of the original test set). (b) "*T-set*", used for recognition studies (55% of the original test set). Since, each speaker has spoken each word only once, the speakers present in the *H-set* of any word are not present in the *T-set* of that word.

The recognition performance of different systems on *T-set* for 8 different sets of 75 words lexicon and one set of 602 words lexicon are given in Table 1. The performance of the 75 word lexicon is the average of the 8 word error rates (obtained for the 8 different sets of 75 words lexicon)..

Table 1: Recognition studies performed on 8 different sets of 75 words lexicon and one set of 602 words lexicon with single pronunciation for each word. Performance is measured in terms of word error rate (WER), expressed in %. Notations: *O*: Auxiliary feature observed, *H*: Auxiliary feature hidden (i.e. integrated over all possible values of auxiliary feature).

Systems		Performance 75 words	Performance 602 words
system-base		4.2	11.0
system-app-p	*O*	2.5	7.3
system-cond-p	*O*	3.5	9.9
	H	4.0	11.3
system-app-e	*O*	5.3	13.3
system-cond-e	*O*	2.9	8.3
	H	3.5	10.2

We extracted the pronunciation variants using the acoustic models of system-app-p for this study, as this system performs better than all the systems and also it better matches the acoustic observation and baseform pronunciation (as observed in last section). For the utterances of each word in *H-set*, we ran the evaluation procedure

1. If it is found that for the majority of utterances of the word the baseform pronunciation is adequate. Then, no pronunciation variants are included.

2. If the above condition is not satisfied, we look for the most frequently inferred pronunciation variant (not diverging far from the baseform) across different utterances, during evaluation. We add it to the lexicon. If there is no commonly inferred pronunciation (it mostly happens for short words.), we extract variants from each utterance such that the confidence level is high and at the same time LD is low.

In the present study, we have done this manually. In future, we would like to automate it using a combined measure such as *comb* described in the last section. The statistics of the test lexicon after adding the pronunciation variants is given in Table 2 (combining all the words in the lexicons of the 8 different test sets).

Table 2: Statistics of test lexicon. The first column mentions the number of pronunciations and the second column gives the number of words with that number of pronunciations.

# of resulting Pronunciation models	Number of words
1	441
2	106
3	48
4	7

We performed recognition studies with the updated lexicon(s). The results of the recognition studies performed on 8 different sets of 75 words lexicon and one set of 602 words lexicon are given in Table 3.

Table 3: Recognition studies performed on 8 different sets of 75 words lexicon and one set of 602 words lexicon with multiple pronunciations . Performance is measured in terms of WER (expressed in %). Notations: *O*: Auxiliary feature observed, *H*: Auxiliary feature hidden. † improvement in the performance is significant compared to the results in Table 1 (with 95% confidence or above)

Systems		Performance 75 words	Performance 602 words
system-base		$3.0^\dagger$	10.1
system-app-p	*O*	$1.7^\dagger$	6.4
system-cond-p	*O*	$2.8^\dagger$	9.2
	H	3.3	10.7
system-app-e	*O*	$4.3^\dagger$	12.0
system-cond-e	*O*	$2.3^\dagger$	7.9
	H	$2.7^\dagger$	9.2

Comparing the performances of the respective systems in Tables 1 and 3, we observe that by adding new pronunciation variants we improve the performance of the 75 words lexicon system significantly. Improvements are also obtained in the case of one set of 602 words lexicon (in some cases an absolute improvement of 1%). This indicates that the addition of new pronunciation variants in the lexicon does not increases the confusion between the words. The interesting point to observe here is that the pronunciation variants were extracted using acoustic models of system-app-p; it could be expected that the extracted pronunciation variants are more suitable for system-app-p as opposed to all the systems. On the contrary, we can see that the addition of pronunciation variants improve the performance of all systems.

7 Summary and Conclusions

In this paper, we proposed an approach based on HMM inference to evaluate the adequacy of pronunciation models. For each lexicon word, the general idea is to start from a *constrained ergodic model*, corresponding to the first-order approximation of the baseform pronunciation and thus only allowing the generation of phonetic sequences basically identical to the baseform pronunciation. This *constrained ergodic model* is then iteratively relaxed to converge towards a fully ergodic HMM, thus allowing all possible phonetic sequences. 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) and (2) the Levenshtein distance with respect to the baseform pronunciation. A good pronunication model should result in a high confidence level and a good stability when relaxing the ergodic HMM. This approach was used to evaluate baseform pronunciation of the words in the lexicon and, extract new pronunciation for the words whose baseform pronunciation were not stable. Recognition studies performed on task-independent speaker-independent isolated word recognition task yielded significant improvement.

In addition to this, in our analytical studies we showed that the proposed pronunciation model evaluation approach can be used to evaluate different acoustic models. When comparing across different acoustic models, we observed that acoustic models trained with both standard features and auxiliary features could improve the stability of the baseform pronunciation of words. This suggests an alternate approach to model pronunciation variation, where the baseform pronunciations can be fixed and the acoustic models are enriched so as to maximize their matching and discrimination properties. This has to be studied further.

In future, we would like to investigate measures to combine the confidence score, Levenshtein distance and "speed of divergence" in order to automatically evaluate the baseform pronunciations and extract pronunciation variants.

8 Acknowledgements

This work was supported by the Swiss National Science Foundation (NSF) under grant MULTI (2000-068231.02/1) and Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2. The NCCR is managed by the Swiss NSF on behalf of the federal authorities. This paper has benefitted from the work of Ait-Aissa Hassou done at IDIAP. The authors would like to thank Guillaume Lathoud and Joanne Moore for their valuable comments and suggestions.

References

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

[BO99] M. Bacchiani and M. Ostendorf. Joint lexicon, acoustic unit inventory and model design. *Speech Communication*, 29(2–4):99–114, 1999.

[MDB01] Mathew Magimai-Doss and Herve Bourlard. Pronunciation models and their evaluation using confidence measures. Technical Report RR-01-29, IDIAP, Martigny, Switzerland, 2001.

[MDSB03] M. Magimai.-Doss, T. A. Stephenson, and H. Bourlard. Using pitch frequency information in speech recognition. In *Eurospeech*, pages 2525–2528, September 2003.

[MJ98] H. Mokbel and D. Jouvet. Derivation of the optimal phonetic transcription set for a word from its acoustic realisation. In *Proceedings of Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 73–78, May 1998.

[Ost99] Mari Ostendorf. Moving beyond the 'Beads-on-a-String' model of speech. In *Proc. IEEE ASRU Workshop*, 1999.

[PFW+95] J. F. Pitrelli, C. Fong, S. H. Wong, J. R. Spitz, and H. C. Leung. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In *ICASSP*, pages 1767–1770, 1995.

[Sar00] Murat Sarclar. *Pronunciation modeling for conversational speech recognition*. PhD dissertation, CSLU, Johns Hopkins University, Baltimore, USA, 2000.

[SC99] Helmer Strik and Catia Cucchiarini. Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29:225–246, 1999.

[SK99] David Sankoff and Joseph Kruskal. *Time Warps, String Edits and Macromolecules: The theory and practise of sequence comparison*. CSLI Publications, Leland Stanford Junior University, 1999.

[SMDB04] T. A. Stephenson, M. Magimai.-Doss, and H. Bourlard. Speech recognition with auxiliary information. *IEEE Trans. Speech and Audio Processing*, 4:189–203, May 2004.

[SO94] Andreas Stolcke and Stephen M. Omohundro. Best-first model merging for hidden Markov models. Technical Report tr-94-003, ICSI, Berkeley, Califronia, USA, January 1994.

[WR99] G. Williams and S. Renals. Confidence measures from local posterior probability estimates. *Computer Speech and Language*, 13:395–411, 1999.