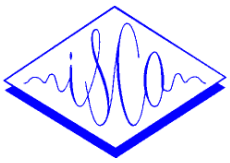

ITRW on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA2002)
Aspen Lodge, Estes Park, Colorado, USA
September 14–15, 2002

IMPLICIT PRONUNCIATION MODELLING IN ASR

Thomas Hain

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
Email: th223@eng.cam.ac.uk

ABSTRACT

Modelling of pronunciation variability is an important part of the acoustic model of a speech recognition system. Good pronunciation models contribute to the robustness and portability of a speech recogniser. Usually pronunciation modelling is associated with the recognition lexicon which allows a direct control of HMM selection. However, in state-of-the-art systems the use of clustering techniques has considerable cross-effects for the dictionary design. Most large vocabulary speech recognition systems make use of a dictionary with multiple possible pronunciation variants per word. In this paper a method for a consistent reduction of the number of pronunciation variants to one pronunciation per word is described. Using the single pronunciation dictionaries similar or better word error rate performance is achieved both on Wall Street Journal and Switchboard data.

1. INTRODUCTION

State-of-the-art automatic speech recognition (ASR) systems are required to operate in complex environments, both acoustically and linguistically. in order to obtain reasonable performance statistical pattern recognition approaches have dominated research in this field of speech recognition over recent decades. The tolerance of stochastic models to misrepresentation has proven to be vital in the development of large scale systems capable of operating in diverse frameworks. The complexity of the task makes a separation into knowledge sources such as acoustic, pronunciation and language models a necessity. For the purpose of the development the assumption of independence of these sources is often used. Recently interest in better integration of all knowledge sources in an ASR system has increased.

The separation of the acoustic models into parts devoted to modelling of certain speech relevant factors is not straight-forward and is often tied to a certain modelling technique. The use of words as recognition units is relatively general and common to very different approaches. The step to recognition of a large number of different words spoken continuously makes this approach infeasible due to lack of training data and due to computational cost. Thus more often words are represented by an abstract representation of the acoustic realisation, the pronunciation string. State-of-the-art medium or large vocabulary ASR systems use pronunciation dictionaries as additional knowledge source to translate the words into model structures. This is especially the case in a Hidden Markov Model (HMM) framework where the generation of word HMMs from phone models is straight-forward.

The use of individual HMMs for phonemes does not provide enough flexibility to model coarticulatory, allophonic effects. Phonemic context can be provided in the form or triphone models while model or state clustering allows to group speech models into atomic speech units. The techniques to capture phone variability are similar in most ASR systems. These techniques are very flexible and allow the construction of powerful classifiers. One side-effect is that the influence of higher level information such as the choice of a particular phoneme for one pronunciation of a word potentially only has a minor effect. In a maximum likelihood framework the construction of models is strongly influenced by the relative frequency of occurrence in the training data. Thus implicitly the use of context allows to the automated construction of almost word specific models. The states in an HMM do not carry a linguistic or phonetic meaning. Even though the left-to-right model is designed to segment speech in time, the matching properties of triphone models are not necessarily tied to good quality in signal segmentation. This is of lesser concern in modelling of clean read speech, but is clearly evident in the highly variable acoustics present in spontaneous or conversational speech [1].

The above properties would suggest that the modelling of pronunciation variation on a symbolic level is problematic as it may be significantly altered by the flexible mappings used on lower levels. The selection of a pronunciation for a word in a recognition dictionary has three effects: determination of the durational range for a particular word; determination of the number of pseudo-stationary states in the word; assignment of the states to other states of other words that may be similar in nature. The durational constraint has potentially only a minor effect on performance while the state selection is also addressed by lower level clustering stages.

The rest of this paper is organised as follows: The next section provides a brief discussion of pronunciation modelling techniques with special focus on implicit modelling techniques. The following section describes a strategy for picking of pronunciation variants to form a single pronunciation dictionary. The experimental sections give details on experiments using single pronunciation dictionaries performed on the Wall Street Journal and Switchboard corpora.

2. MODELLING OF PRONUNCIATION VARIATION

The factors influencing the realisation, transmission and perception of a speech signal are manifold. A separation of the speech signal from acoustic channel effects is only possible to a limited degree. Even though it is difficult to draw strict boundaries between pronunciation modelling and general acoustic modelling [2], pronunciation modelling most often is concerned with the definition, selection and model representation of symbols that can be used to describe the acoustic realisation of an utterance. Explicit information is used in knowledge-based approaches that are often combined with statistical learning algorithms to model general

variability (see e.g [2] for a detailed discussion).

Most ASR systems use multiple pronunciations per word in training and test dictionaries to model pronunciation variation. More recently the progress towards the recognition of spontaneous speech has triggered increased interest in pronunciation modelling. The reason for the attention is the astonishingly poor performance of speech recognisers in these scenarios, which to a significant degree can be attributed to a substantial increase of pronunciation variability[3]. Nevertheless the improvements reported for extended pronunciation models do not yield the anticipated performance [4].

2.1. Pronunciation variants and networks

The freedom to use of multiple different pronunciations for a particular word can be represented either by a list of possibilities or, more compactly by compressing the list into a network. Whereas theoretically a list representation does have a potentially trivial network representation, the use approximations in decoding may show undesired effects. In any case the use of multiple representations for a particular word increases the confusability with other words as the distance in pronunciation between words usually becomes smaller. Thus the benefit from adding new variants has to be balanced with added confusability. Most pronunciation dictionaries used in large vocabulary ASR systems are generated by automatic rule based systems and corrected manually. This process is expensive, especially if the dictionary has to be constructed from scratch. The dictionary used in this paper is based on the LIMSI American English 1993 WSJ dictionary [5]. All pronunciations for words added to the original dictionary have been checked manually.

One way to control the relationship between pronunciation variants is by using pronunciation probabilities in addition to acoustic and language models. Depending on the approach probabilities may come out of a variant generation process or may be estimated from data. Since data sparsity is an issue, smoothing of the probability estimates is important. One simple but effective technique is to obtain the pronunciation probability estimates from alignment of the training data [6].

2.2. Implicit modelling

Explicit modelling of pronunciation variability has a series of disadvantages. Knowledge based approaches are usually expensive and do not always translate well to new domains. Secondly explicit information necessarily is only available on a high level which implies a coarse influence on the model structure. In conjunction with data-driven methods, data sparsity is a problem due to the low symbolic rate. If the targeted pronunciation effects are not of primary importance, even explicit methods are difficult to assess.

Another option is to model the variability with statistical models that do not need to operate on a high level or require an interpretation of the underlying hidden data. If these models are tied in with the standard HMM framework, joint training can yield improved performance. The disadvantage of such schemes is that specific targeting of pronunciation effects is difficult and that other acoustic effects can have an effect on performance.

HMMs usually make use of mixtures of distributions of the exponential family. Assuming that pronunciation variation is sufficiently modelled by a substitution of stationary states, the use of mixture distributions would be sufficient to model the variation. However other effects, for example gender dependence, can be equally modelled in that framework. Data sparsity and longer-span contextual effects may make the use of higher level information desirable.

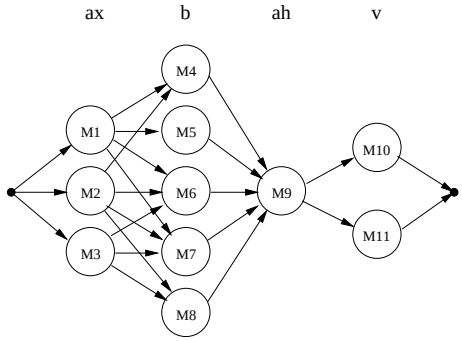

Fig. 1. Basic HMS-HMM structure for modelling of substitution effects. The nodes M_i represent HMMs with arbitrary topology .

The following briefly discusses some options for implicit pronunciation modelling. However, an exhaustive discussion is beyond the scope of this paper.

2.2.1. Parameter tying

If pronunciation effects are representable by symbolic replacements, the automated tying of parameters can be used to capture and accurately model pronunciation variability. One particular advantage of this approach is that maximum likelihood or discriminative criteria can be used to assess the quality of clusters. Examples are the phonetic decision tree clustering of states or complete models[7], or the use of tied mixture models[8]. Other more complex parameter tying schemes such as fenones[9] have a more direct relationship to pronunciation strings. More recently [10] has introduced a method for the soft-tying of states, which was used by [3] to model pronunciation variability in spontaneous speech. It is important to note that parameter tying schemes are usually capable of modelling substitution effects, but maintain the temporal structure of the encapsulating HMMs. More flexibility can be achieved by optional deletions and insertions in HMM selection [11].

2.2.2. Hidden model sequences

Hidden model sequence HMMs (HMS-HMMs) [12] provide an extension to parameter tying. The basic idea is to replace the deterministic mapping provided by phonetic decision trees with a stochastic model. This model provides the mapping between a given sequence of phonemes and a particular realisation as HMM sequence. In principle arbitrary mappings between the two sequences can be trained within an Expectation-Maximisation framework. In order to limit the confusability one particular realisation of a HMS-HMM is designed to model only substitutions. In this case a set of HMMs for each phoneme is defined. Each model represents a possible realisation of that particular phoneme. However, dependent on the phonemic context a probability is assigned to each model. Fig. 1 shows the basic network structure for a word model. In that case the contextual probabilistic dependency on neighbouring models is ignored. It is important to note that an additional controlled sharing of models between phonemes al-

lows a detailed modelling of pronunciation effects. For a detailed discussion of HMS-HMMs see [13].

2.2.3. Alternatives

In recent years the use of HMMs for speech recognition has attracted much criticism. Apart from the many theoretical shortcomings in parameter estimation the main objection addresses the fact that HMMs do not capture the underlying, hidden processes that drive speech production and pronunciation. One of the main focal points is the modelling of articulator movement (e.g. [14]). Another important approach is the use of generalised linear Gaussian models [15] to capture the dynamic aspects of the speech signals. In these cases hidden data is assumed to control the realisation of the speech signals. Either the hidden data is un-observable by definition or has no physical or linguistic interpretation, or cannot be obtained easily.

3. SINGLE PRONUNCIATION DICTIONARIES

The modelling of pronunciation variability in HMM based speech recognisers cannot be separated from other acoustic modelling issues. The use of context and more specifically the structural aspects of HMM sets such as parameter tying or mixture modelling have a significant impact on the performance of each component.

The use of multiple pronunciation variants in dictionaries is commonly assessed on the basis of an existing single pronunciation dictionary. In these cases the addition of new and obvious variants usually brings improvements as long as the confusability is kept low. The addition of variants is based on the assumption that the starting point, the initial dictionary was optimal. However, without knowledge of the task it is difficult to assess the anticipated confusability a priori. The selection of one pronunciation variant has an influence on all other words in the recognition dictionary. Consistency in phonetic transcription may be of greater importance an improved representation of the training utterances. Taking this into account the inverse problem is of interest: Given a well performing multiple pronunciation dictionary, is it possible to derive a consistent and well performing single pronunciation dictionary, and which cross-effects can be perceived in conjunction with other pronunciation modelling techniques?

In experiments the LIMSI 1993 WSJ multiple pronunciation (MPron) dictionary[5] serves as reference. Under the assumption that dictionaries may be task dependent, single pronunciation (SPron) dictionaries are specifically constructed using training data for the particular task. The following procedure was adopted for the selection of pronunciations from the baseline MPron dictionary:

1. Pronunciation variant frequency

   By Viterbi-aligning the transcriptions of the training data, the frequency of each pronunciation in the training dictionary is obtained.

2. Initial dictionary

   The pronunciations for each word in the baseline MPron dictionary are sorted according to frequency of occurrence in the training data. If a word was observed in the training data, any associated unseen pronunciation variants are deleted.

3. Merging of phoneme substitutions

   For a given word each pair or pronunciation variants is aligned using DP alignment. If phonemes are only substituted, the variant with the higher frequency of occurrence is retained and the frequency of the second variant is added. If the variants occurred equally often the selection is random.

After these stages the variants remaining in the dictionary fall into two categories: Variants observed in the training data but which cannot be solely described by phoneme substitutions; and variants for words not observed in the training data. For both cases a statistical model may be trained to serve in decision making. In order to assess the importance of this selection process two different approaches, further denoted as methods *F* and *P*, have been implemented. In method *P* the decision process is based on training of a simple statistical model: Given two phoneme strings a and b we would like to determine if a is a realisation of a source pronunciation $\mathbf{s}$ and b is the associated target pronunciation $\mathbf{t}$ or vice versa. This decision can be made by comparing the probabilities of these events:

$$P(\mathbf{s}=a,\mathbf{t}=b) \lessgtr P(\mathbf{s}=b,\mathbf{t}=a) \tag{1}$$

This equation can be simplified using Bayes rule:

$$P(\mathbf{t}=b|\mathbf{s}=a)P(\mathbf{s}=a) \lessgtr P(\mathbf{t}=a|\mathbf{s}=b)P(\mathbf{s}=b) \tag{2}$$

Under the assumption that the priors for the source transcription are equal, i.e. $P(\mathbf{s}=a^{(i)}) = P(\mathbf{s}=a^{(j)})$ for any pair (i,j), they priors can be discarded in the decision process. The use of the chain rule can provide an estimate for $P(\mathbf{t}|\mathbf{s})$:

$$\hat{P}(\mathbf{t}|\mathbf{s}) = \prod_{i=1}^{M} P(t_i|\mathbf{t}_1^{i-1},\mathbf{s}) \tag{3}$$

Assuming that the sequences are DP-aligned a simple model for computing Eq. 3 is given by:

$$\hat{P}_P(\mathbf{t}|\mathbf{s}) = \prod_{i=1}^{M} \hat{P}(t_i|s_i) = \prod_{i=1}^{M} \frac{N(t_i;s_i)}{N(s_i)} \tag{4}$$

where $N(t_i;s_i)$ is the frequency that symbol s_i is source and symbol t_i is target. Note that $N(t_i;s_i) \neq N(s_i;t_i)$. The counts can be obtained from the pronunciation variants observed in the training data (using step 3). Consequently the final steps of method *P* are:

4. Model estimation

   Obtain frequencies for Eq. 4 from pronunciation variants with counts. In order to smooth the probability estimates one is added to all counts.

5P. Variant selection

   For all remaining variants select on the basis of Eq. 1. Note that dealing with insertions and deletions simply involves the introduction of an addition symbol to the standard set of phonemes.

An even simpler approach is to use an approximate solution for Eq. 4. Since the basic decision rule is unaltered when using log-probabilities, Eq. 4 can be modified accordingly. Given two specific sequences a, b the decision rule Eq. 1 can be rewritten as

$$C_a +_i \sum_{i=1}^{M} \log N(a_i; b_i) \lesseqgtr C_b + \sum_{i=1}^{M} \log N(b_i; a_i)$$

where C_a and C_b are constants representing the sum of the log of the frequencies of the phonemes in the training data. If we assume that $C_a \approx C_b$ and use $\log x \approx -1 + x$, a direct comparison of counts as obtained in step 4. can be made.

$$\sum_{i=1}^{M} N(a_i; b_i) \lesseqgtr_i \sum_{i=1}^{M} N(b_i; a_i) \quad (5)$$

Method *F* only uses Eq. 5 for substitution only pronunciation pairs:

5F. Variant selection

If pronunciations associated with a word not observed in the training data can be aligned with substitutions only, Eq. 5 is used for the decision.

6F. Random selection

Any remaining pronunciation variants are either selected by frequency, if available, or on a random basis.

The use of method *F* is only feasible if a good coverage of the test vocabulary is present in the training data.

4. EXPERIMENTS

State-of-the-art ASR systems ideally should be capable to operate in diverse environments. Whereas initially the focus was on the transcription of read speech, more recently interest is devoted to recognition of spontaneous or conversational speech from obtained from a variety of acoustic channels. Word error rates (WERs) on conversational speech are significantly higher than those obtained on read speech. Read speech may seem as ideal test-bed for pronunciation modelling as it lacks other acoustic distortions. However, the natural environment of for example a telephone conversation allows for a much greater variability in pronunciation which is substantially more difficult to model. The following sections describe experiments conducted on WSJ as a sample of a read speech corpus, and Switchboard, as example for transcription of conversational telephone speech.

4.1. Wall Street Journal

The ARPA 1994 Hub1 unlimited vocabulary NAB News development and evaluation test sets [16] were used for experiments on this corpus. Experiments used gender independent triphone mixture-Gaussian tied-state HMM models with 12 mixture components for each speech state. The models are similar to the one used in [17]. The speech signal is represented by a stream of perceptual linear prediction cepstral coefficients derived from a mel-scale filterbank (MF-PLP). A total of 13 coefficients, including c_0, and their first and second order derivatives were used. This yields a 39 dimensional feature vector. Cepstral mean normalisation was performed on a per sentence basis. The acoustic training data consisted of 36493 sentences from the SI-284 WSJ0 and WSJ1 sets.

The dictionaries used in training and test are based on the 1993 LIMSI WSJ lexicon [5] which contains multiple pronunciation variants per word. For training of the baseline model set

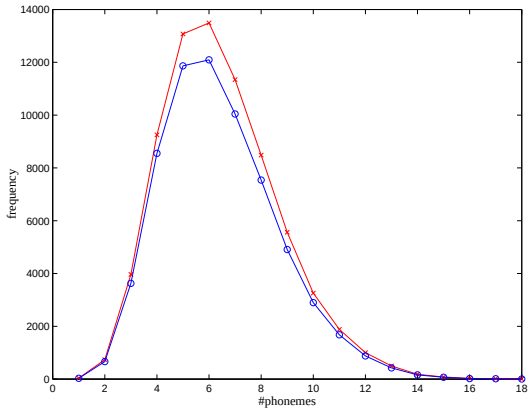

Fig. 2. Number of pronunciations as a function of pronunciation variant length. The graph shows the distribution for the MPron(x) and SPron1(o) dictionaries.

Models	Dict	H1 Dev	H1 Eval	Average
Baseline	MPron	8.97	9.65	9.33
Baseline	SPron1	10.95	11.97	11.48
ReEst	SPron1	9.37	10.31	9.86
ReEst	MPron	9.07	9.50	9.30

Table 1. %WERs on the WSJ H1 development and evaluation test sets. Results are obtained by rescoring trigram lattices. All models are are state-clustered 12 mixture triphone models.

an MPron dictionary containing 13665 words was used. On average the dictionary contained 1.18 pronunciations variants per word and the maximum number of variants per word was 8. The MPron dictionary used for recognition tests was considerably larger with 65466 entries and an average of 1.11 pronunciations per word. A maximum of 12 variants per word was used. However, most of the 6779 words with variants had only one alternative pronunciation. The difference in the average number of variants per word between training and test dictionaries illustrates that more variants are used for frequent words.

Three single pronunciation dictionaries were constructed using different strategies: SPron1 used pronunciation variant frequencies both from the WSJ training set as well as from the Switchboard corpus (see Section 4.2) and method *P* for variant selection; SPron2 was built using frequencies obtained from WSJ training data only and method *P* for pronunciation variant selection; for SPron3 a purely random variant selection process was adopted. Figure 2 shows the histograms over the length of pronunciations for the MPron and SPron1 test dictionaries. Note that after automatic variant selection the pronunciation length distribution is for the two dictionaries remain similar.

A first series of experiments was conducted to explore the interaction between model structure and pronunciation dictionary. Table 1 shows word error rates obtained by rescoring of trigram lattices for various combinations of model sets and dictionaries. The *Baseline* model set is trained using the MPron dictionary, while *ReEst* denotes four iterations of Baum-Welch re-estimation steps of the *Baseline* models using the SPron1 dictionaries in train-

Dict	#states	H1 Dev	H1 Eval	Average
MPron	6447	8.97	9.65	9.33
SPron1	6419	9.05	9.95	9.53
SPron2	6425	9.33	9.93	9.64
SPron3	6486	9.65	10.95	10.24

Table 2. %WER results on the WSJ H1 Dev and eval test sets using different dictionaries for both training and test. #states denotes the number of clustered states in the model set.

		H1 Dev	H1 Eval	Average
HMM	Mpron	8.97	9.65	9.33
HMS-HMM	MPron	9.08	9.15	9.12
HMM	SPron1	9.05	9.95	9.53
HMS-HMM	SPron1	8.65	9.43	9.06

Table 3. %WER results on the WSJ H1 Dev and eval test sets.

ing. The use of the SPron1 dictionary in test only yields a severe degradation in performance which can be only partially regained by re-estimation of model parameters with the SPron1 training dictionary. Note that the use of the MPron dictionary on the re–estimated model set still gives the same performance.

Table 2 shows results obtained with HMM sets trained with four different dictionaries. In all cases model training involved re-clustering as well as successive mixture splitting. The parameters in state clustering were chosen to obtain models with a comparable number of model parameters. Not surprisingly the poorest word error rate performance is obtained using the SPron3 dictionary. The difference between SPron1 and SPron2 is minor and both show slightly poorer performance compared to the MPron baseline. Note that in the SPron1 case the training from the start showed significantly better results than that obtained by re-estimation of model parameters only (in Table 1).

The word error rate performance using single pronunciation dictionaries is slightly worse than that for the MPron dictionary. Table 3 shows experimental results using HMS-HMMs to model pronunciation variation as described in Section 2.2.2. Whereas only a moderate improvement is achieved on the basis of the MPron dictionary, similar performance can be achieved by using the SPron1 setup [1].

4.2. Switchboard

The Switchboard corpus is a large collection of conversational telephone speech. Due to the nature of the data pronunciation modelling on this task attracted widespread attention (e.g. [18]). A detailed description of the techniques and models used in the transcription of Switchboard data would go beyond the scope of this paper. For a description of techniques as well as training and test sets see [19].

Compared to the situation on WSJ, training and test dictionaries have a larger overlap. The simple method *F* was used for the construction of the SPron dictionaries in training and test. The test dictionary as used contains 54598 words with an average 1.10

[1]Note that all HMS-HMM sets are initialised with the corresponding standard HMM models and thus have the same number of HMM parameters.

Dict	PProb	Swbd1	Swbd2	Swbd2cell	Average
MPron		22.4	39.4	39.0	33.5
SPron		21.6	37.9	37.8	32.3
MPron	×	21.5	37.9	38.1	32.4
SPron	×	21.3	37.7	37.4	32.0

Table 4. %WERs obtained by rescoring 4-gram (fgintcat) lattices of the dev01sub test using models trained on the h5train02 training set (HLDA, VTLN, 28 mixture components). Models were trained using ML estimation of parameters.

Dict	PProb	Swbd1	Swbd2	Swbd2cell	Average
MPron		19.3	35.6	35.8	30.1
SPron		19.4	35.2	35.1	29.8
MPron	×	19.1	35.0	35.6	29.8
SPron	×	19.6	34.9	34.9	29.7

Table 5. %WERs obtained by rescoring 4-gram (fgintcat) lattices of the dev01sub test using models trained on the h5train02 training set (VTLN,HLDA, 28 mixture components). Models were obtained using a Minimum Phone Error Rate training criterion.

pronunciations per word. The training dictionary consists of 34651 words with an average of 1.14 variants per word. Table 4 shows word error results on the 2.97 hour dev01sub test set. This test set covers data from three different subsets of the Switchboard corpus, Switchboard1 (Swbd1), Switchboard2-Phase3 (Swbd2) and data collected via mobile phones, Switchboard2-Phase4 (Swbd2cell). Note the different levels of difficulty for each subset. As can be observed, the use of a SPron dictionary yields lower word error rates on all subsets. The table shows a further comparison with the use of pronunciation probabilities (see Section 2.1). Note that the CU-HTK systems silence models are treated as part of the pronunciation. The use of pronunciation probabilities with a SPron dictionary denotes word specific probabilities for the silence variants. The overall performance of the SPron and the MPron models is similar, apart from the Swbd2cell set.

Previously all models were trained using standard Baum-Welch re-estimation. Table 5 shows the corresponding experiments using models trained with the discriminative MPE criterion [19]. While word error rates using the SPron dictionaries are poorer on the Swbd1 part of the test set, a good result on the difficult Swbd2cell data remains.

Similar to experiments conducted on the WSJ corpus the use of SPron dictionaries was investigated in conjunction with HMS-HMMs. For this purpose a smaller part of the Switchboard 1 corpus, the 18 hour MiniTrain set was used for model training. Tests were conducted on two 30 minute Switchboard test sets (MTtest and WS96DevSub). The HMM baseline system has 2954 states and 12 mixture components. Table 6 shows WER results obtained by rescoring of trigram lattices. Whereas the performance of MPron and SPron models is very similar, a slight advantage is given in the SPron case.

	Dict	MTtest+WS96devsub
HMM	MPron	45.04
HMM	SPron	44.89
HMS-HMM	MPron	43.38
HMS-HMM	SPron	43.12

Table 6. %WER results on Switchboard using models trained and tested with dictionaries containing one (SPron) or multiple (MPron) pronunciations.

5. CONCLUSIONS

A method for constructing a dictionary with only one pronunciation entry per word from a good reference dictionary was presented. The performance of this dictionary in terms of word error rates was investigated on both a read and a conversational speech corpus. In both cases the use of the proposed dictionary gave comparable or better performance than the standard baseline system using multiple pronunciation variants. This suggests that implicit modelling of pronunciation variation can perform equally or better than explicit a-priori knowledge when using a good starting point. Furthermore, due to the side-effects with other techniques it can be beneficial to use a single pronunciation dictionary in order to allow other techniques to learn the necessary model structures. This has been demonstrated by using HMS-HMMs to model substitution variability with consistent results on both corpora.

6. ACKNOWLEDGEMENTS

The author would like to thank BBN for providing the MiniTrain training and MTtest test set definitions and IBM for an SUR equipment award. This work is in part supported by the DARPA EARS project.

7. REFERENCES

[1] S. Greenberg, "The Switchboard transcription project," 1996 LVCSR summer workshop technical reports, Center for Language and Speech Processing, Johns Hopkins University, 1996, `http://www.icsi.berkeley.edu/real/stp`.

[2] H. Strik and C. Cucchiarini, "Modelling pronunciation variation for ASR: A survey of the literature," *Speech Communication*, vol. 29, pp. 225–246, 1999.

[3] M. Saraçlar, H. J. Nock, and S. Khudanpur, "Pronunciation modelling by sharing Gaussian densities across phonetic models," *Computer Speech and Language*, vol. 14, pp. 137–160, 2000.

[4] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect of speaking style on LVCSR performance," in *Proc. of ICSLP'96*, 1996, pp. S16–S19.

[5] J.-L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker, "The LIMSI Nov93 WSJ system," in *Proc. 1994 ARPA Spoken Language Technology Workshop*, Plainsboro, NJ, Mar. 1994, pp. 125–128.

[6] T. Hain, P. C. Woodland, G. Evermann, and D. Povey, "New features in the cu-htk system for transcription of conversational telephone speech," in *Proc. of ICASSP'01*, 2001, pp. 57–60.

[7] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. 1994 ARPA Human Language Technology Workshop*. 1994, pp. 307–312, Morgan Kaufmann.

[8] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Trans. ASSP*, vol. 38, no. 12, pp. 2033–2045, Dec. 1990.

[9] L. R. Bahl, J. R. Bellegarda, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "A new class of fenonic Markov models for large vocabulary continuous speech recognition," in *Proc. of ICASSP'91*, 1991, vol. 1, pp. 177–200.

[10] X. Luo and F. Jelinek, "Probabilistic classification of HMM states for large vocabulary continuous speech recognition," in *Proc. of ICASSP'99*, Apr. 1999, pp. 2044–2047.

[11] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Lolje, J. McDonough, H. J. Nock, M. Saraçlar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand–labelled phonetic corpora," *Speech Communication*, vol. 29, pp. 209–224, 1999.

[12] T. Hain and P. C. Woodland, "Dynamic HMM selection for continuous speech recognition," in *Proc. of EUROSPEECH'99*, Sept. 1999, vol. 3, pp. 1327–1330.

[13] Ph.D. thesis.

[14] Li Deng, "Speech recognition using the autosegmental representation of phonological units with the interface to the the trended HMM," *Speech Communication*, vol. 23, pp. 211–222, 1997.

[15] A.-V. I. Rosti and M. J. F. Gales, "Generalised linear gaussian models," Tech. Rep. CUED/F-INFENG/TR420, Cambridge University Engineering Department, 2001.

[16] D.S. Pallett, J. G. Fiscus, W.M. Fisher, J. S. Garofolo, B. A. Lund, and A. Martin, "1994 benchmark tests for the ARPA spoken language program," in *Proc. ARPA Workshop on Spoken Language Systems Technology*, 1995, pp. 3–5.

[17] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "The development of the 1994 HTK large vocabulary speech recognition system," in *Proc. ARPA Workshop on Spoken Language Systems Technology*, 1995, pp. 104–105.

[18] W. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. J. Nock, M. Riley, M. Saraçlar, C. Wooters, and G. Zavaliagkos, "Pronunciation modelling using a hand-labelled corpus for conversational speech recognition," in *Proc. of ICASSP'98*, 1998, vol. 1, pp. 313–316.

[19] P. C. Woodland, G. Evermann, M. J.F. Gales, T. Hain, A. Liu, G. Moore, D. Povey, and L. Wang, "CU-HTK APRIL 2002 SWITCHBOARD SYSTEM," Rich Transcription Workshop, Vienna, VA, 2002.