Proc. of ASRU 2003, Dec 2003, U.S Virgin Islands

IMPROVEMENTS IN ENGLISH ASR FOR THE MALACH PROJECT USING SYLLABLE-CENTRIC MODELS

Abhinav Sethy, Bhuvana Ramabhadran

Human Language Technologies
IBM T. J. Watson Research Center
Yorktown Heights, NY.

sethy@usc.edu, bhuvana@us.ibm.com

Shrikanth Narayanan

Speech Analysis and Interpretation Lab
Integrated Media Systems Center
Department of Electrical Engineering
University of Southern California
shri@sipi.usc.edu

ABSTRACT

LVCSR systems have traditionally used phones as the basic acoustic unit for recognition. Syllable and other longer length units provide an efficient means for modeling long-term temporal dependencies in speech that are difficult to capture in a phone based recognition framework. However, it is well known that longer duration units suffer from training data sparsity problems since a large number of units in the lexicon will have little or no acoustic training data. Previous research has shown that syllable-based modeling provides improvements over word internal systems, but performance has lagged behind crossword context-dependent systems. In this paper, we describe a syllable-centric approach to English LVCSR for the MALACH (Multilingual Access to Large spoken ArCHives) project. The combined modeling of syllables and context-dependent phones provides a 0.5% absolute improvement in recognition accuracy over the state-of-the-art cross word system for the heavily accented and spontaneous speech seen in oral history archives. More importantly, we report on the importance of the improved recognition of names and concepts that is crucial for subsequent search and retrieval.

1. INTRODUCTION

Automatic speech recognition (ASR) systems typically focus on short-time information distributed over periods of 10-20 ms. A speech signal is parameterized by partitioning it into overlapping frames of 20-30 ms; several of these consecutive frames are spliced together and projected to a lower-dimensional space using a linear discriminant feature space transform to ensure maximum phonetic discriminability. This feature space representation coupled with decision trees that capture the immediate phonetic context of these feature vectors and search constraints from the language model are used to decode the best matching word or phone sequence for the given speech signal. Such short-term representations of speech have proven to be successful in a wide range of recognition tasks. However, there are good indications [1][5][6] that capturing information distributed over longer periods of time, such as syllabic or word level time span, can lead to substantial gains in recognition accuracy.

The number of different acoustic units required for a given recognition task is a function of the vocabulary size and the nature of the underlying acoustic units. For phonemes the number of basic models (without context modeling) is fixed for a given language. However, when using syllable or word size units, the number increases in general with the vocabulary size. Many of these units are pronunciations of words which are not used frequently and will have poor coverage in the training data. Sparsity of training data has been the main hindering block in using longer acoustic units for large vocabulary speech recognition tasks (LVCSR). For small vocabulary tasks such as alphabet or digit recognition, longer units (typically word level units) have been used successfully. The training data sparsity problem can be partially addressed by combining context dependent phone and syllabic units in a single framework. Nevertheless, it is still hard to build context-dependent syllables (obtain lengthier contextual information) or to train phones with syllables as context. In this paper, we address these issues by designing techniques to obtain the appropriate syllable internal context, word internal context and cross word context for the phones in a mixed phonetic syllabic system. As we will describe in our analysis of experimental results (Sections 6 and 7), context is a very important factor in the design of the syllable system . We also describe the use of skip states in HMM topologies to limit the effects of poorly trained syllabic states.

The best performing ASR system is one that uses competing phonetic and mixed syllabic-phonetic paths in parallel. In addition, the use of a penalty for the syllabic path for functional monosyllabic words is described. This approach compensates for the lack of syllabic context resulting from the trade-off between data sparsity and context-independent syllables which causes the mixture distributions to have softer variances. Our system was trained and evaluated on the MALACH [2] corpus. The next section describes the MALACH corpus in detail. The motivation for syllable based LVCSR is described in Section 3. Section 4 describes the design of the syllabic lexicon. Section 5 describes the baseline crossword context-dependent system and the training setup. Sections 6 and 7 describe the different syllable modeling approaches, their effect on performance and an analysis of the errors. This paper concludes with some insights into the usefulness of syllable-centric modeling, especially in the context of MALACH.

2. SYLLABLES AND MALACH

MALACH, (Multilingual Access to Large Spoken Archives), is an ongoing effort that aims to achieve a quantum leap in our ability to access the contents of large, multilingual, spoken archives by advancing the state of the art in automated speech recognition (ASR), information retrieval (IR) and other component technolo-

gies, by utilizing the world's largest digital archive of video oral histories collected by VHF[1]. VHF was created to record the first-hand accounts of Holocaust survivors, liberators, rescuers and witnesses and disseminate that information to future generations [2]. The MALACH corpus consists of unconstrained, natural speech filled with disfluencies, heavy accents, age-related coarticulations, uncued speaker and language switching and emotional speech collected in the form of interviews from over 52000 speakers in 32 languages. Approximately 25000 of these testimonies are in English, spanning a wide range of accents, such as Hungarian, Polish, Yiddish, German, Italian, French, Czech, Hebrew, Ukrainian etc. A good number of words uttered in this corpus are foreign words or sequences of words spoken in a foreign language, unfamiliar names and places. The corpus consists of elderly speech, where the age of the interviewees range from 56 years to 90 years. The age-related coarticulation effects (natural deletion of phones) contribute significantly to the high word error rates seen in this corpus.

From a syllable LVCSR point of view, MALACH is an interesting and challenging corpus with an extraordinary amount of heavily accented and co-articulated speech. Unlike spontaneous speech corpora like Switchboard, MALACH has a richer syllabic content which accentuates the training sparsity problem. For SWB 90% of the training data can be covered with just 800 syllables [6]. For MALACH the number of syllables needed for this level of coverage of acoustic data is more than 1800. In addition, the currently available training data for the English portion of the MALACH corpus is limited to around 65 hours [4] which makes the data sparsity problem more acute for syllables. Nevertheless, the heavily accented nature of this corpus makes a natural testbed for evaluating the performance improvements obtained through acoustics-based pronunciation modeling of longer length units such as syllables. The state-of-the-art English ASR system for the MALACH project is described in detail in [4].

3. OUR APPROACH TO SYLLABLE MODELING IN LVCSR

The use of an acoustic unit with a longer duration facilitates exploitation of temporal and spectral variations simultaneously. Parametric trajectories and multi path HMMs [7][8] are examples of techniques that can exploit the longer acoustic context, and yet have had marginal impact on phone-based systems. Units of syllabic duration or longer are much more effective in capturing the cross phone correlations and temporal dependencies. In this paper, we present recognition systems which use a combination of syllable and phone units. The motivation for using syllables as basic acoustic units along with phonemes for large vocabulary transcription tasks comes from recent research on syllable based recognition [5], [6], as well as studies of human perception [9], [10] which demonstrate the central role that the syllable plays in human perception and generation of speech. One important factor that supports the use of syllables as an acoustic unit for recognition is the relative insulation of the syllable from pronunciation variations arising from addition and deletions of phonemes as well as coarticulation. In studies of the Switchboard corpus [11] it has been shown that syllables have a deletion rate of 1% whereas the deletion rate for phonemes is 12%. Robustness to pronunciation variations is particularly important for the heavily accented speech in the MALACH corpus. The major challenge in using syllables and other longer length units for recognition is the training data sparsity problem. In [6] this problem is partially resolved by using only those syllables which have good coverage in the acoustic data. However, the syllable being a larger unit, requires more training data than phone sized units and hence proper training of syllable level models using flat initialization strategies, as described in [6] is difficult. We propose the use of context-dependent HMM state-level (leaf) alignments to train syllable state models (Section 6). Furthermore, syllable models which do not have enough acoustic data coverage need to be replaced by their corresponding phonemes in the lexicon. We describe the lexicon design process in the next section.

[1]VHF, or The Survivors of the Shoah Visual History Foundation.

Word Pronunciation	Number of words (%)	Occurrence in training (%)
Pure phonetic	20	49
Pure Syllabic	22	38
Mixed	56	11

Table 1. Coverage of words having purely phonetic, syllabic and mixed phonetic syllabic pronunciations

4. SYLLABLE LEXICON

The first step in designing a syllable based recognition system is to prepare the syllabic lexicon. We represent syllables in terms of the underlying phone sequence. Thus given a phonetic transcription of the speech in a standardized format like Wordbet or IPA we can write a syllable representation by coming up with a set of syllable symbols from the phonemes comprising the syllable. For example, *Ghetto* with the phonetic transcription 'G EH T OW' can be represented in syllabic terms as 'G_EH' 'T_OW'.

The next stage in designing a syllable lexicon is to identify the phone clusters, which correspond to the correct syllabic representation. The process of clustering phones to get a syllable representation is called syllabification. Syllabification principles are described in [13] as a set of rules which define permitted syllable-initial consonant clusters, syllable-final consonant clusters and prohibited onsets. Syllabification software available from NIST [14] implements these rules and comes up with a set of alternative possible syllable clusters given a phoneme sequence which are used to generate the syllabic lexicon.

Acoustic data severely restricts the range of syllable models that can be trained. Figures 1 and 2 show the coverage in the acoustic data for the different syllable models. In both figures, the X-axis represents the minimum number of times a syllable occurs in the acoustic data. In Figure 1 the Y-axis represents the coverage of acoustic data that can be achieved for a given minimum occurrence count and in Figure 2 it represents the number of syllables which have at least that many occurrences. This distribution is important while deciding on the lexical representation of words in the dictionary. The distribution for the MALACH corpus is representative of a typical conversational corpus like SWITCHBOARD or TIMIT but is more restrictive because of more limited training data and a more varied syllabary.

The syllable lexicon was built in the following manner. Based on the coverage distributions illustrated in Figures 1, and 2, a threshold value for the occupancy counts was set to determine the syllables for which models will be built. A word was then repre-

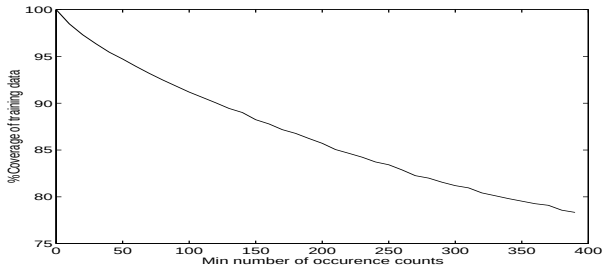

Fig. 1. Percentage of acoustic data covered by syllables with a given minimum occurrence count

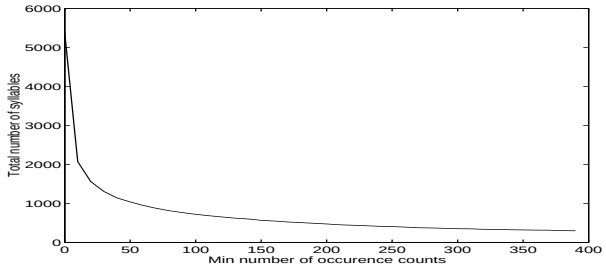

Fig. 2. Number of syllables with a given minimum occurrence count

sented in the dictionary as a complete phonetic sequence, complete syllabic sequence or a mixture of phones and syllables. Table 1 shows the percentage of pure phonetic words, pure syllabic words and mixed words in terms of both the vocabulary coverage and actual training data coverage which is the weighted sum of the number of times a particular word occurs in training data. However, this threshold does not truly reflect the number of feature vectors that will be aligned to the different syllable states. This can lead to undertrained states even though the syllable might have a large occupancy count. Hence, the potential for many more syllables to be dropped at a later stage when building syllable models exists. This issue is described in more detail in section 6. We begin with a description of the state-of-the-art crossword context dependent phonetic system that is used for generating initial alignments for the syllable models.

5. MALACH SYSTEM DESCRIPTION

The English ASR system uses acoustic models constructed using 65 hours of English interviews from 260 speakers in the VHF corpus. The compressed audio signal from the MPEG1 videos is down-sampled to 16KHz; 24-dimensional Mel frequency cepstral coefficients (MFCC) and 60-dimensional transformed features are then extracted. The 60,000 word lexicon built from existing cataloging information and a study of the frequency of occurrence of uncommon words, has good coverage of names and places likely to be mentioned during interviews. The language model was built by interpolating the 1.7M words from the MALACH corpus with data from Broadcast News (50M words) and Switchboard (3M words) corpora. Pronunciations for the many unseen words in this corpus were derived using existing dictionaries and tools using spelling-to-sound rules. In addition to the speaker independent models, we also built speaker adaptive models on this corpus (SAT) using a constrained maximum-likelihood linear regressing (fMLLR) [4] feature space transform on the features. Context-dependent HMM states were obtained by querying the crossword phonetic context using a decision tree. The data at the leaf nodes of the tree were modeled with diagonal Gaussian distributions via a BIC-based procedure [4] and trained using multiple iterations of the EM algorithm. The test corpus consists of 30-minute segments of interviews from 30 randomly selected speakers. The development test set used throughout this paper is an hour of data from 20 speakers, selected from this test corpus [4]. This context-dependent phonetic system (LC) comprised of 58,000 Gaussian distributions.

6. EXPERIMENTS AND RESULTS

The syllable system was bootstrapped from the crossword left-context (LC) phonetic system described in the previous section. The additional number of syllables caused an increase in the number of Gaussians to 100,000 inclusive of the 58,000 Gaussians from the LC system. To bootstrap the syllable models, state level alignments obtained using the LC system were used. Thus for every word in the training transcriptions the phonetic pronunciation was used to seed the syllables. Since a syllable topology matches the state sequence of the phones constituting the syllable, a state by state alignment can be generated for the syllabic system directly from the phonetic alignments. Diagonal Gaussian distributions were built for the data aligning to each of the syllable states using the BIC criterion for determining the number of Gaussians and a few iterations of EM. Figure 3 illustrates the average number of Gaussian distributions per syllabic state and Figure 4 the minimum number of distributions that some states get modeled by.

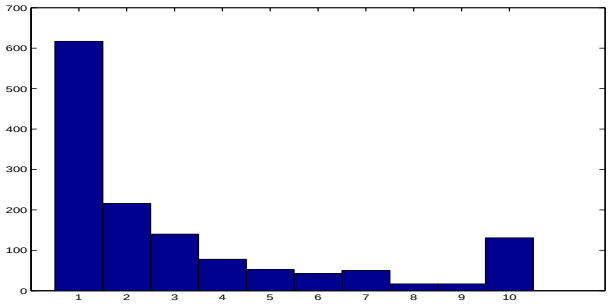

Fig. 3. Average number of Gaussians per state; the X-axis is truncated to a maximum of 10 Gaussians

Our analysis of the errors made by the above system indicated that many errors could be attributed to syllables which had a good occurrence count but because of the skew in state level alignments i.e., some states have very few feature vectors aligned to them, a single Gaussian was assigned to some states. We allowed such undertrained states to be skipped based on the unreliability of the distributions modeling these states. An example of such a topology is given in Figure 5 for the syllable AY_DD which has six states.

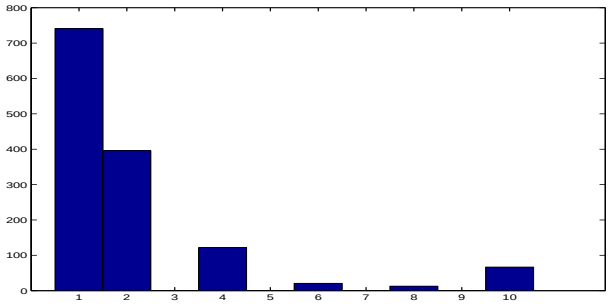

Fig. 4. Minimum number of Gaussians per syllabic model; the X-axis is truncated to a maximum of 10 Gaussians

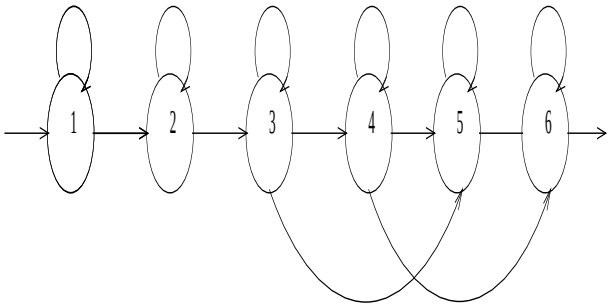

Fig. 5. Skip topology for the syllable AY_DD(EYED).

After EM training states 4 and 5 had only one Gaussian. With the introduction of skip arcs, the syllable system could correctly recognize many syllables which had some undertrained states. However, the overall word error rate degraded with skip arcs (See Table 4). Analysis of the errors of this system show that the insertion rate with skips was much higher. This can be attributed to the shortened minimum duration topology resulting because of skips. Many of these syllables were then deleted from the syllabary.

In the developmental stages we used a small set of 255 utterances to evaluate the different strategies that we were experimenting with. We first compared models trained with syllable internal [1], word internal [6] and cross word context for just the phones (Table 2). For the word internal system the phonetic pronunciation of the word was used to traverse the left-context crossword phonetic decision tree. The resulting leaves of the tree were then used as basic units of pronunciation in the lexicon. For the crossword context systems the context for a phone model was defined by the phonetic sequence corresponding to the syllables which were in that phone's neighboring context. For example, the context for the phone 'TD' in 'AX B_AO_R TD' is the phonetic sequence 'AX B AO R'. Thus, in effect, the cross word context phonetic decision tree was copied along with the context-independent syllabic states. In all cases the syllable models were context free. The performance of the syllable internal and word internal systems was significantly behind the baseline crossword LC system.

Given the nature of errors made by the syllabic system and words that can be corrected by the syllabic system, we built a system which had a mixed syllabic-phonetic variant and a pure phonetic variant for every pronunciation in the dictionary. For example, consider the word 'MAMMOTH'. The word has a phonetic pronunciation 'M AE M AX TH'. The syllabic representation was 'M_AE M_AX_TH'. However the syllable 'M_AX_TH' has very little training data (below the selected threshold) and the corresponding syllable model is not trained. Thus the pronunciation of this word in the mixed syllabic-phonetic system is 'M_AE M AX TH'. The dual pronunciation system (DP) then includes both the mixed syllabic-phonetic pronunciation and the pure phonetic variant of the same pronunciation. This system effectively chooses the maximum likelihood from the syllable or the corresponding phonetic variant. This is necessary in view of the data sparsity and the context free nature of syllable models. The motivation behind such an approach is to retain the gains from syllable modeling and a pure phonetic approach which does not suffer from data sparsity issues and is context dependent. Table 3 shows the comparison in performance of the single syllabic pronunciation system and the dual pronunciation system. Having dual pronunciations allows the model to choose between the two paths based on their likelihood scores, thus resulting in better performance. While the mixed pronunciation path captures the syllabic context within a word well, the pure phonetic system has better trained acoustic models with no data sparsity issues and compensates for any crossword context and undertrained models from the mixed pronunciation path.

System Type	WER (%)
Syllable Internal context	41.2
Word Internal context	39.8
Crossword context	39.2
Baseline cross word phonetic	32.3

Table 2. Performance comparison of syllable systems having different phonetic contexts with the baseline crossword context dependent phonetic system for single syllabic pronunciations

Pronunciation variants	WER (%)
Single	39.2
Dual	33.2

Table 3. Word Error Rate (WER) for syllabic systems having syllabic pronunciations and dual pronunciations

7. ERROR ANALYSIS

To measure the effectiveness of syllable models we calculated the number of times a syllable occurred in a utterance and the number of words that were corrected in that utterance. This would test the hypothesis that the syllable models help in getting better time

Topology	WER (%)	Insertion Rate (%)
Skip	41.2	9.0
No skip	39.2	7.4

Table 4. Word Error Rate (WER) for skip and non skip HMM topologies

Error Type (%)	Dual Pronunciations with Penalty	Baseline: LC Phonetic System	Syllabic Pron. only
Word Error Rate	30.5	31.0	36.3
Percentage correct	74.4	73.7	70.1
Insertions	4.9	4.8	6.4
Deletions	4.8	5.1	5.3
Substitutions	20.8	21.1	24.6

Table 5. Comparison of the dual pronunciation system with penalties and the combined syllabic-phonetic system with the baseline crossword context-dependent LC system

marks for the phonetic boundaries aiding the overall alignment process. Our results confirmed the hypothesis in two ways. For many utterances we discovered that even though the decoder chose the syllabic representation for only one word, multiple neighboring words in the hypothesis would get corrected. Secondly, the syllable system helped improve the spotting of names and longer words (See Table 6). These words usually have mixed syllabic-phonetic pronunciation and the improved performance can be attributed to the syllabic sections of the pronunciation since the phonetic path is identical to the baseline LC system. In the MALACH project this is particularly important for the search and retrieval of segments of speech relevant to the mention of a name, place or a concept.

An analysis of the errors of the dual pronunciation system indicates that most of the errors could be attributed to a higher insertion rate for short functional monosyllabic words such as 'the', 'my', 'in' etc. The pronunciation of these words changes significantly with the word context (especially observed in heavily accented speech) and since the syllable models are context free and have larger variances, they are unable to correctly model these words in context. To solve this problem we decided to insert a penalty on the syllabic path for these words which improved the performance significantly. We believe that if we had sufficient data to estimate many of these syllables, we could build context-dependent syllable models and this would eliminate some of the insertion errors caused by the syllable models corresponding to these monosyllabic words to some degree. The monosyllabic penalty is applied by a Finite State Transducer composition on the rescored lattice. Table 5 lists the gains obtained with and without penalizing the monosyllabic words on our evaluation test set. This test set is a subset of the test set described in [4] and contains about 5 hours of speech from 55 speakers (about 30K words). We also experimented with modeling the monosyllabic words separately; this reduced the data from the multi-syllabic words of which these syllables are a part of even further, leading to under-estimated models and poorer performance.

Table 6 reflects the kind of errors that are corrected by the syllable system. The first example illustrates how the two paths in the mixed syllabic system work in tandem to produce an overall improvement in recognition accuracy. The words 'buy' and the first occurrence of 'today' were recognized correctly through the syllabic path, while the phonetic path scored higher for the second utterance of 'today'.

The third and fourth examples confirm our initial hypothesis that the syllabic models perform better with recognition of names and places. 'Prague' and 'Moiche' are two words that are important for retrieval and word spotting that are correctly recognized via the syllabic path. In the final example, the correct recognition of the word 'in' by the syllabic models caused the rest of the utterance to be recognized correctly as well. This is particularly important as search on the word 'Amsterdam' would now yield a result which was otherwise impossible to do.

On analyzing the types of errors that the syllable system could recover, we found that most recoveries came from long content words where the syllable was embedded in the middle surrounded by phones. A plausible explanation for this is that the crossword context dependent nature of that word pronunciation was accounted for by the phones which were at the word pronunciation boundaries whereas the word internal pronunciation variations were captured to a great degree by the syllable. At the time of writing this paper, experiments were still under way with a second-pass reestimation of the syllable models. Results using these models will be presented at the workshop.

8. CONCLUSION

In this paper, we have proposed a syllabic-centric approach to acoustic modeling that is additive to the results obtained with context-dependent crossword phonetic models. We have described the design, implementation and other practical issues with syllable based LVCSR, specifically in the context of the MALACH corpus. Our best system is a mixed syllabic-phonetic system with two pronunciation variants per lexical entry in the dictionary. This system performs better than the best crossword phonetic system to date for English ASR for the MALACH project. Our results indicate that syllables and potentially lengthier acoustic units can help in improving recognition accuracy for LVCSR. However, in order to benefit maximally, careful attention has to be paid to the data sparsity and context issues. To the best of our knowledge, we are unaware of any syllable-based LVCSR research reported in the literature that has provided additive gains to a cross-word context dependent system. The next section offers some insights into improving the modeling of such units further.

9. FUTURE WORK

This paper illustrates the potential improvements to recognition accuracy that can be realized through the syllabic-centric approach to acoustic modeling. The dual pronunciation system produces a lesser number of deletion and substitution errors compared to the baseline system. As a next step, we plan to explore context-dependent syllabic models to alleviate many of the data sparsity problems when incorporating contextual information. Our acoustic models were trained on 65 hours of transcribed data that was available at the time of this work. We now have 200 hours of transcribed speech to work with. We also plan to automatically derive the best representation of pronunciations in a lexicon using a combination of syllabic and phonetic units that discriminates well across words that share subsets of these units. Our initial syllabic classification work with parametric (polynomial) trajectory mixture models as a means to model the temporal evolution of features for variable length syllabic units provided encouraging results. An analysis of the errors made by the syllable system showed that many words can be correctly recognized if they can be discriminated from other similar sounding syllables. N_IY_Z and L_IY_Z

Reference	and it was what we buy today ready in the market this this today it's
Mixed Syl-Ph System	and it was what we buy today ready in the market this this today it's
Context-Dep. Ph System	and it was what we buy today ready in the market JUST DOES NOT today it's
Reference	there was another typical thing when i was a child
Mixed Syl-Ph System	there was another typical thing when i was a child
Context-Dep. Ph System	there *** another typical thing when i was a *****
Reference	to transfer MONIES from prague to budapest
Mixed Syl-Ph System	to transfer MONEY from prague to budapest
Context-Dep. Ph System	to transfer MONEY from BLOCK to budapest
Reference	his name was ******* MOICHE KIRSHENBAUM
Mixed Syl-Ph System	his name was MOISCHE CROATIAN BOMB
Context-Dep. Ph System	his name was MOLLY SHE CROATIAN BOMB
Reference	THAT this was DONE IN in our city in sighet ** MARAMARIS
Mixed Syl-Ph System	AND this was **** COMMON in our city in sighet MY MORRIS
Context-Dep. Ph System	AND this was **** COMMON in our city HASIDIC MY MORRIS
Reference	and HER father was well known in amsterdam in the textile
Mixed Syl-Ph System	and MY father was well known in amsterdam in the textile
Context-Dep. Ph System	and MY father was well known AND STAND in the textile

Table 6. Examples of words recognized correctly and incorrectly by the syllable system

are examples of such confusable pairs. Also, the number of parameters needed to model a syllable with polynomial trajectory mixtures will be far fewer than that needed by a conventional HMM-based model as the trajectory representation is tied to the order of the polynomial. For completeness, we plan to evaluate the performance of our syllable-centric modeling on other transcription tasks, such as the Switchboard Evaluation task, voicemail messages, etc. Finally, we plan to generalize this framework to longer length acoustic units, i.e., speech segments which may not coincide with phone or syllable boundaries. This framework would support a lexicon representation in terms of these units which may be modeled using different families of distributions.

10. ACKNOWLEDGMENTS

This project is part of an on-going effort funded by NSF under the Information Technology Research (ITR) program, NSF IIS Award No. 0122466. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

11. REFERENCES

[1] Abhinav Sethy, Shrikanth Narayanan, "Split-Lexicon based hierarchical recognition of speech using syllable and word level acoustic units", *ICASSP*, Hong Kong, April 2003, vol.1, pp. 772-776

[2] MALACH website: http://www.clsp.jhu.edu/research/MALACH

[3] http://www.isip.msstate.edu/projects/switchboard/doc

[4] Bhuvana Ramabhadran, Jing Huang, Michael Picheny, "Towards Automatic Transcription of Large Spoken Archives - English ASR For The Malach Project", *ICASSP*, vol. 1, pp. 216-220, April 2003

[5] Abhinav Sethy, Shrikanth Narayanan and S. Parthasarthy, "A syllable based approach for improved recognition of spoken names", *Proceedings of the ISCA Pronunciation Modeling Workshop*, Estes Park, Colorado, September 2002

[6] Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington and J. Picone, "Syllable-Based Large Vocabulary Continuous Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 358-366, May 2001.

[7] H. Gish and K.Ng, "Parameter trajectory models for speech recognition", *Proceedings of ICSLP, Philadelphia. PA, 1996*, pp 466-469.

[8] F.Kormazskiy, "Generalized mixture of HMM's for continuous speech recognition" , *Proceedings of ICASSP, Munich, Germany, 1997*, pp 1443-1446.

[9] Kirchhoff, K., "Syllable-level desynchronisation of phonetic features for speech recognition", *International Conference of Spoken Language Processing 1996*, pp 2274-2276.

[10] Su-Lin Wu, Brian Kingsbury, Nelson Morgan, and Steven Greenberg, "Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition", *ICASSP-98*, Seattle, pp. 721-724.

[11] S. Greenberg, "Speaking in Shorthand - A Syllable-Centric Perspective for Understanding Pronunciation Variation", *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, The Netherlands, May 3-6, 1998.

[12] Odell J, Ollason D, Woodland P, Young S , Jansen J, "The HTK Book for HTK V2.0", Cambridge University Press, Cambridge, UK, 1995.

[13] D. Kahn, "Syllable-Based Generalizations in English Phonology", Indiana University Linguistics Club, Bloomington, Indiana, USA, 1976.

[14] W.M. Fisher, "Syllabification Software", *http://www.itl.nist.gov/div894/894.01/slp.htm*, The Spoken Natural Language Processing Group, National Institute of Standards and Technology, Gaithersburg, Maryland, U.S.A., June 1997.