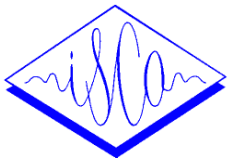
ISCA Archive
http://www.isca-speech.org/archive



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

A SYLLABLE BASED APPROACH FOR IMPROVED RECOGNITION OF SPOKEN NAMES

Abhinav Sethy, Shrikanth Narayanan

Integrated Media Systems Center
Department of Electrical Engineering -Systems
University of Southern California
[sethy,shri] @sipi.usc.edu

S. Parthasarthy

AT&T Labs-Research
sps@research.att.com

ABSTRACT

Recognition of spoken names is a challenging task for speech recognition systems because of the large variations in speaking styles, linguistic origins and pronunciation found in names. The complex linguistic nature of names makes it difficult to automatically generate pronunciation variations. For many applications the list of names tends to be in the order of several hundred thousands, making spoken name recognition a high perplexity task. Use of multiple pronunciations to account for the variations in names further increases the perplexity of the recognition system substantially. In this paper we propose the use of the syllable as the acoustic unit for spoken name recognition and show how pronunciation variation modeling with syllables can help in improving recognition performance and reducing the system perplexity. We present results comparing systems which use context dependent phones with syllable based systems, and demonstrate that a significant increase in recognition accuracy and speed, can be achieved by using the syllable as the acoustic unit for spoken name recognition. With a finite state grammar network for spoken name recognition, the observed recognition error rate for the syllable-based system was 40% less than the phone-based system. For syllable bigram based information retrieval schemes the observed recognition error rate was about 60% less than the corresponding phone system.

1. INTRODUCTION

Spoken name recognition is a key component of speech recognition applications. There is an increased interest in this problem with the recent decision by the LVCSR community to adopt the named entity task as the next step towards a speech-understanding framework. A typical application for spoken name recognition is in directory assistance or name dialing systems[1][2]. Other applications include city name recognition as part of a travel system, caller identification for banks etc. For most applications the list of names tends to be in the order of hundreds of thousands, making the spoken name recognition task a high complexity problem. In addition, the large variability in name pronunciation, both at the segmental and suprasegmental level, significantly decreases recognition accuracy. Names have multiple valid pronunciations that evolve as a product of various socio-linguistic phenomena. Specifically in a country like USA with a broad cultural base, there is a considerable variability in linguistic origins of names. A large number of names have foreign origin and depending on the speaker's linguistic background, they are pronounced differently. As an example, the name *Abhinav* which has an Indian (Sanskrit) origin is typically pronounced as 'ae b hh ih n ae v' by a native speaker of American English whereas a native speaker from India pronounces it as 'aa b hh ih n eh v'. To achieve good recognition accuracy a large number of the possible pronunciation variations need to be included in the dictionary.

Automatic pronunciation generation techniques based on neural nets or tree based approaches exist to generate the different possible pronunciations of a given word. However these techniques require a large set of words and their different pronunciations for training[3][4]. In addition performance of such schemes is limited for names which have non-native origins since generating valid pronunciations requires an understanding of the original language and its phonology. Embedding this knowledge into an automatic pronunciation generation system is not easy and thus, we often require manual augmentation of the names pronunciation dictionary. It should be also noted that variations in 'non native' pronunciation include stress placement or prosody variations which are very difficult to cover in a dictionary in a consistent manner. The problem of pronunciation variability is tied to the language model perplexity problem. Inclusion of multiple pronunciations increases the perplexity of the underlying language (grammar) network substantially, since the recognizer has to consider multiple paths for every name.

To address pronunciation and speaker variability in a better way we explore the usefulness of a larger acoustic unit, the syllable. The syllable is a basic unit of speech consisting of two or more phones, including a nucleus that is usually a vowel, and is generally perceived as having no interrupting pause within it. In English we can categorize different types of syllables by their consonant(C) and vowel (V) content. The typical syllable is a CVC syllable i.e., a consonant pair with a vowel between them. An example of this kind of syllable is 't eh n' corresponding to the word *ten*. The syllable is defined based on human speech perception and speech production phenomena, typically augmented by stress patterns. The syllable provides a promising framework for improving the spoken name recognition accuracy without the use of multiple pronunciations (Sec.3). To address the high complexity problem of the spoken name recognition task we propose a syllable centeric information retrieval technique which allows for significant improvements in speed with a minor tradeoff in accuracy.

In this paper we will compare the performance of a syllable based recognition system designed along the lines of [5] with systems based on context dependent phones and a hybrid system which combines syllable and phone based units. In the next section we describe the basic architecture of our name recognition systems. In section 3 we discuss the motivation for using syllables for spoken name recognition and describe the design of the

syllable-based recognizer in section 4. Training strategies and corpora are described in section 5. In section 6 we present the comparative performance evaluation results and discuss our findings. In the concluding section, we provide a summary of our work, the major findings and an outline for future research.

2. SPOKEN NAME RECOGNITION SYSTEMS

The standard approach for spoken name recognition is to use a Finite State Grammar (FSG) based recognition network, in which all the required names along with their possible pronunciations are taken as arcs or alternate paths for evaluation. The recognizer matches the input utterance against all possible names and their variations and selects the name, which matches best. One could use unigram or bigram name (word) level statistics to weight the different paths. However it is difficult to obtain these statistics from real usage data e.g. directory applications. Considering this we gave equal weight to all the names, although it is possible to improve the recognition performance for names which are very common (such as John, Smith) by giving them a higher weight. As is evident from the design, the perplexity of FSG recognition networks can be prohibitive for very large name lists which can be in the order of 150K or more words for many directory applications.

For the spoken name recognition tasks, we observed that as the perplexity of the FSG recognition network grows the phone level recognition accuracy drops and for very large namelists it becomes comparable to the accuracy of a bigram based phone recognition system. Based on this observation, a promising approach (Figure 1) for cases with large word lists is to use inverse dictionary lookup techniques to recognize the name. That is we identify the underlying N best phone sequences based on a n-gram (phone) language model and then use statistical string matching to find out the best candidates from the name list using the dictionary. The statistical string matching process compares the phone sequence with the pronunciations for the different names in the name list and selects names which have similar pronunciation. This N-best list (or equivalent lattice) can be rescored in a more constrained way. For example, an FSG recognition network can be generated for rescoring. This can be seen as an information retrieval problem. The advantage of such a scheme is that it makes possible a substantial reduction in computational complexity with a small tradeoff in accuracy. Performance of such techniques depends on the accuracy of the n-gram based recognizer and the pronunciation variations covered in the dictionary for name retrieval. Our experimental results (see section 6) revealed that the phone is not a good unit for information retrieval based name recognition schemes. This can be attributed primarily to the differences in speaking style, pronunciation variations and the high rate of phone addition and deletion in natural speech. In addition the limited context information that can be embedded in phone level units reduces the accuracy of the recognizer.

3. SYLLABLES FOR SPOKEN NAME RECOGNITION: MOTIVATION

To improve the accuracy of FSG based or reverse lookup based spoken name recognition systems we propose the use of the syllable as the acoustic unit for name recognition. As discussed in the previous section, a phone based information retrieval scheme for name recognition gives very low accuracy, since the n-gram

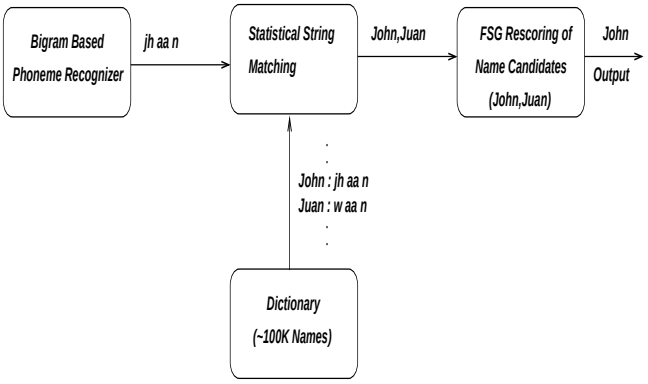


Fig. 1. Information retrieval scheme for name recognition provides scalability

phonontactic grammar recognizer has very low accuracy (around 40-50%). The high recognition accuracy of a loop based syllable recognizer allows us to use information retrieval ideas and hence acheive significant reduction in complexity with a small accuracy tradeoff.

The syllable presents an acoustic unit which because of its extended context information and close relation with human perception of speech[11], is well suited for dealing with differences in pronunciation, which are common in names. The longer duration of a syllable, which spans multiple phones, implies that pronunciations differing in a single phone (over a syllabic time span) are less likely to affect the syllabic sequence recognition. In addition to linguistic background differences, significant changes in pronunciation are bought about by phenomena like coarticulation and deletion of phones, which are very common in casual speech. The syllable provides a natural framework for integrating coarticulatory phenomena. In studies on the Switchboard corpus [6][7] it has been shown that syllables had a deletion rate of 1% percent whereas the deletion rate for phones was 12%. The relative robustness of the syllable against small changes in pronunciation allows us to capture the phone level variations in a single syllabic representation. Thus we are able to use a single representation for names in a syllable FSG network unlike phone FSG networks which need to be augmented by rule based pronunciation generation techniques followed by extensive manual verification.

In addition, the syllable provides a convenient framework for incorporating suprasegmental prosodic information into the recognition system. Incorporation of such information in phone based system can be acheived through techniques like parameter trajectories or multipath HMMs [8][9]. However these techniques have been only marginally successful so far. Recent research on stochastic modeling of phones demonstrates that recognition performance can be improved by exploiting correlations in temporal and spectral structure, which can be done more effectively at the syllabic level than at the phone level.

The syllable can be used for the Information Retrieval centered scheme outlined in section 2, with the help of a syllable-based dictionary(syllabry). The underlying syllable sequence is identified using a unigram or bigram based syllable recognizer and then statistical string matching in the syllabry is used to identify the reduced name candidate list for FSG rescoring. This outperforms similar phone based systems as the dictionary has lesser valid instances (for the same name) to match against (which makes it eas-

ier to classify a particular sequence). In addition, the embedded context information of the syllable implies that the underlying sequence would be identified more accurately than the phone case.

4. SYLLABLE BASED SPOKEN NAME RECOGNITION: DESIGN

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 Worldbet or IPA we can write a syllable representation by coming up with a set of syllable symbols from the phones comprising the syllable for eg. *Junior* with the phonetic transcription 'jh uw n y er' can be represented in syllabic terms as 'jh_uw' 'n_y_er'.

The next stage in designing a syllable lexicon is to identify the phone clustering, which corresponds to the correct syllabic representation. The process of clustering phones to get a syllable representation is called syllabification. Syllabification principles are described in [12] as a set of rules which define permitted syllable-initial consonant clusters, syllable-final consonant clusters and prohibited onsets. Syllabification software available from NIST [13] implements these rules and comes up with a set of alternative possible syllable clusters given a phone sequence. These alternatives differ in terms of rate or degree of casualness, informality, or lack of monitoring. In addition, the NIST software allows for ambisyllabicity, that is, a consonant may simultaneously be the final segment of one syllable and the initial segment of the next. Thus the word *bitter* becomes 'b_iy_t' 't_er' with the phone 't' being shared by two different syllables. However these were very infrequent in the names dataset that we are interested in. In order to keep the complexity of the syllable system low we decided to restrict the syllable dictionary to one entry per name. Thus we ignore the ambisyllabic information. Accuracy of the syllabification process can be improved by using stress information which was not available in the name pronunciation dictionaries we used.

The phone level HMM models have the same basic topology with equal number of states, for different phones. However syllable models require different number of states depending on their size. A syllable comprising four phones such as 's_w_eh_l' requires more number of states than single phones or other shorter syllables such as 't_eh_n'. To account for this the number of states was chosen to be three times the number of phones comprising the syllable. This scheme of having a fixed number of states in the syllable model for every phone makes it easy to initialize the syllable models from phone level models. Other schemes for deciding the number of states for syllable models are under investigation. To cover the high contextual variance in a syllable length unit, we chose to have 16 Gaussian mixtures for every state in the syllable model.

Unlike the phone-based system, we decided not to use context models for the syllable system. The number of syllables in our datasets are of the order 10^3 and introducing context dependencies would bring in a combinatorial explosion by increasing the number of possible symbols to a few million. Also the spoken name recognition task typically involves recognizing one or two words (first and/or lastname). In our view, incidences of firstname/lastname and middlename are not based on established linguistic or grammar principles and can be misleading for building cross-word context models. Thus we are restricted to just in-word contexts, which would lead to poor training of the context models since number of syllables in a typical name is small.

To initialize the models for the syllable recognizer we use pre-trained Context Dependent (CD) phone models. The number of states in a given syllable model is the sum of the number of states of the constituent phone models. Moving from the leftmost phone,we pick the initial state parameters from the corresponding CD phone models. As an illustration consider the syllable 'b_iy_t'. Assuming 3 states per phone, states 1-3 in the syllable model will be initialized using the CD phone b-iy, states 4-6 from the CD phone b-iy+t and states 7-9 from the CD phone iy-t. Thus we need to first build a CD phone recognition system for seeding the syllable recognizer.

The syllable recognizer design is complicated by the large number of syllables possible in a language like English. It is difficult to cover all such syllables even with a phonetically balanced speech corpus like TIMIT. It is important to have a fallback arrangement for the recognition system, so that it can capture speech content for which the syllable-based recognizer has no acoustic units in the training data. This problem was addressed in two ways.

The first method restricts itself to only syllable units. Since the complete name list for recognition is available, it is possible to make models for all the possible syllables in the lexicon using the Context Dependent phone recognizer units. All the syllable models in the lexicon are initialized by CD phone models as described earlier. The training data covers only a certain number of the syllables in the lexicon. If sufficient training data is available for a given unit, the corresponding model gets updated. Models for which sufficient training data is not available are not modified. We will refer to this setup as the (pure) syllable recognizer.

The other technique we used for addressing the syllable training problem is to use a hybrid recognizer similar to [5]. The syllable units which have good coverage in the training data are used along with phone level units. Syllables with poor coverage (i.e. only a few occurrences in the training data) are replaced by their phone level representation in the lexicon. The context for phone models was taken to be the last phone in the left syllable along with the first phone in the right syllable. As an illustrative example, consider the phone level representation 'ax n d r uw'. The syllabic representation would be 'ax_n_d r_uw'. The syllable 'ax_n_d' (corresponding to *and*) would be covered adequately in the training data. Thus we split the lexicon entry as 'ax_n_d r uw'. The context for the phone model 'r' is d-r+uw and for 'uw' it is r-uw. No context information is used for the unit 'ax_n_d'.

5. TRAINING: CORPORA AND IMPLEMENTATION

As the first step we built three recognition systems (phone, syllable and hybrid) using the TIMIT speech corpus. For the TIMIT corpus, we had 1900 syllables with about 70% of the words being either monosyllabic or bisyllabic.

The speech data from TIMIT was downsampled to 8 kHz. 26 mel frequency cepstral coefficients were extracted at a frame rate of 10ms using a 16ms Hamming window. First and second order differentials plus an energy component were used. For the baseline phone based recognizer, 46 three-state left-to-right phone models were initialized and trained on hand labeled data provided in the TIMIT corpus. These were then cloned to yield triphone level models, which underwent reestimation. Tree based clustering was used for state tying to ensure proper training of the models. Output distributions were approximated by four Gaussians per state.

Subsequently the syllable and the hybrid system were initialized as described in section 4 and were trained on the acoustic data. We used 16 Gaussian mixture models per state for the syllable-based system to allow for its larger acoustic size and contextual variance. After training we performed preliminary testing to check the accuracy of the recognizers and fine-tune the parameters on TIMIT.

The primary speech corpus of interest to us for spoken name recognition is the OGI NAMES corpus[14]. The NAMES corpus is a collection of name utterances, covering first, last and full names, collected from several thousand different speakers over the telephone. The name pronunciation is fairly natural since the speakers are not reading the names off a list. Word level transcriptions are provided for all name utterances and some of the utterances are also labeled phonetically. The phonetically labeled files were used to make a names dictionary, which was augmented with some additional name entries from public domain dictionaries like Cambridge university BEEP dictionary and the CMU dictionary. The names corpus is sampled at 8Khz and has about 6.3 hours of speech data. There are about 10000 unique names in the corpus and it covers 40% of the bigram phonetic contexts possible. Tables 1 and 2 describe the occurrence frequency for words of different syllabic count for the TIMIT and NAMES corpus. As can be seen, most names are bi or tri syllabic unlike TIMIT, which has a higher monosyllabic content mainly due to functional words such as 'and', 'the'. Also, words with smaller syllable count are used more frequently in generic sentences of the nature found in TIMIT [5].

We used the models trained on TIMIT to bootstrap the models for the NAMES database. Both the TIMIT and the NAMES dictionaries were merged to yield a single phonetic dictionary, which was then converted to a syllabic dictionary. For the phone level recognizer we used the context independent phone models from TIMIT as initial prototypes and used them to build a NAMES CD phone system. For the syllable and hybrid system we used the final TIMIT models as the prototypes for the syllables common between the two databases, which are around 1200 in number. Table 3 shows the distribution of the syllables common between TIMIT and the NAMES database for different syllable lengths. We can see that a large number of the shorter syllables (2-3 phone length) can be initialized from TIMIT. The remaining syllables in the NAMES lexicon were initialized using the techniques described in section 4 from the NAMES CD phone models.

6. RESULTS

Comparative performance evaluation between the various syllable based and phone based systems is presented for both the FSG network based spoken name recognition and the information retrieval scheme. As discussed in section 4, the syllable recognizer can be initialized in two different ways. The first scheme which provides full coverage of the syllables in the lexicon, will be refered to as the syllable recognizer. The alternate design strategy in which we restrict syllable units to those which have adequate coverage in the training data will be refered to as the hybrid recognizer. The initialization and training of all models was done in the manner described in section 4 and 5 using both the TIMIT and NAMES databases.

Number of syllables	1	2	3	4	5
Words	23%	47%	23%	5%	0.8%

Table 1. Distribution of words and their syllable count for the TIMIT corpus. Total number of words was 8900.

Number of syllables	1	2	3	4	5
Words	13%	50%	30%	6%	0.3%

Table 2. Distribution of words and their syllable count for the NAMES corpus. Total number of words was 10000.

Syllable Length	2	3	4	5
Number of common syllables	30%	53%	15%	2%

Table 3. Distribution of syllables common to TIMIT and NAMES and their length. Total number of common syllables is around 1200.

Recognizer Type	Name recognition Accuracy (%)
Context Independent Phone Recognizer	45
Context Dependent Phone Recognizer	63
Context Independent Hybrid Recognizer	75
Context Independent Syllable Recognizer	80

Table 4. Recognition rate for different FSG based spoken name recognition systems

6.1. Scheme I: FSG Networks

As a first step, we did comparative evaluation of the phone based recognizer with the syllable and the hybrid recognizer for the FSG case. In this experiment the training vocabulary for recognition completely covers the list of names for recognition. FSG based name recognition is useful when the name lists are small and correspondingly the recognizer perplexity is low. A section of the Names database comprising 6000 utterances was used for evaluating the recognition performance and the remaining 4000 utterances in the NAMES database were used for training. We compared the performance of the three recognition systems on this set. The results are given in Table 4. The results for phone models compare well with previous results on similar size name lists[15]. As can be seen from these results for the FSG recognition task, the performance of the two syllable-based recognizers is significantly better than the phone based recognizers.

We next compared how the performance of the phone based and the (pure) syllable based recognizer scales with increasing word list size. We trained both the systems on 1K names and increased the size of the test name list from 1K to 10K. The results are shown in Figure 2 which shows the recognition accuracy with increasing vocabulary size. As can be seen from the figure the (rate of) accuracy drop in the syllable based system is less than the drop for the phone-based system.

6.2. Scheme II : Information Retrieval

In the case of very large word lists we use n-gram based models to identify the underlying unit sequence (phones or syllables) and then use reverse dictionary lookup based on statistical string matching to identify the name (see section 2). We used a bigram model for phones as well as syllables and measured the accuracy with which both the systems find out the underlying unit sequence. The bigram models were trained from both the TIMIT and the NAMES data. Syllable based recognizer was able to give a unit recognition rate of 63% whereas the phone recognizer accuracy was 45%. Since the performance of statistical string matching techniques is critically dependent on the accuracy with which the underlying units are correctly recognized, we expect the syllable-based system to outperform the phone based system for information retrieval schemes.

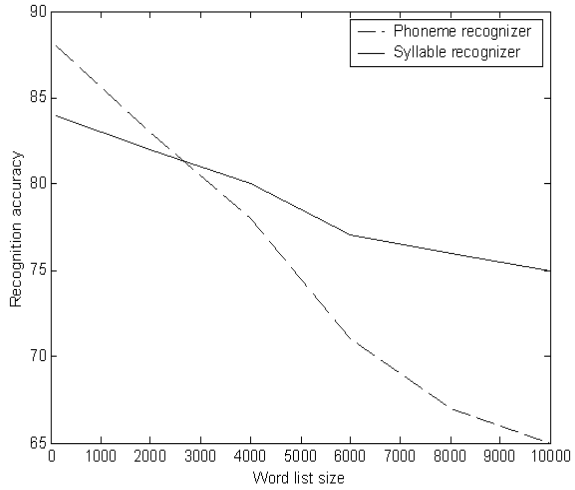


Fig. 2. Plot of recognizer accuracy vs word list size

To verify this we designed reverse lookup based recognition systems using syllable as well as phone models. In the first stage we identify the best 2 underlying unit sequence candidates (based on a bigram language model) for both the systems. In the next stage a name candidate list was selected using reverse lookup in the dictionary. The reverse lookup was based on the Levenshtein distance crieterion which is defined as the minimum number of characters that must be added, removed, or changed in order to transform one string into another. For every name in the name list the corresponding phone sequence was compared with the loop based recognizer output using the Levenshtein distance. All names whose distance was less than a threshold were selected for creating the FSG network for the second stage. The threshold for selection was taken to be a function of the recognized sequence length. For the syllable recognizer we split the recognizer output sequence to the corresponding phoneme sequence in order to calculate the Levenshtein distance.

Our experimental results, which cover a range of spoken name recognition tasks show that a syllable-centered approach outperforms phone-based system. The high unit sequence recognition rate of syllables is of special interest since it allows us to use Information retrieval techniques for spoken name recognition, which leads to a large reduction in the system complexity.

Acoustic unit	Name recognition accuracy after FSG rescoring (%)
Context Independent Phone	45
Context Dependent Phone	61
Context Independent Syllable	73

Table 5. Spoken name recognition accuracy for the information retrieval scheme after FSG rescoring of compacted name list

7. CONCLUSION

Spoken name recognition is a challenging task because of the large variations in speaking style and pronunciation found in names. Modeling the pronunciation variations by a dictionary containing multiple pronunciations is a difficult task considering the potential variability in the linguistic nature of names. In addition, inclusion of multiple pronunciations in a recognition network leads to a substantial increase in the system perplexity. In this paper we have described an alternate technique for pronunciation modeling for spoken name recognition that is based on the use of the syllable as the basic acoustic unit. The syllable-based system gives substantial performance improvements in terms of recognition accuracy over context dependent phone based schemes that are typically used. Performance analysis of the scheme indicates that increasing recognition word list size has lesser impact on the recognition accuracy of the syllable system than on phone based systems. Also the performance of recognition systems, which use n-gram based language models to prune search space, is better if syllable is used as acoustic unit.

A critical issue with larger units such as syllables is that they are not adequately covered in the training data. At this stage we have experimented with two different schemes for the syllable recognizer. In the first scheme we initialized models for all the possible syllables in the lexicon using the Context Dependent phone recognizer units. Thus we had models for all syllables in lexicon even though there was no acoustic data to further train the syllable models. In the second scheme we used phone units in combination with the syllable units which were adequately covered in the training data. We plan to investigate other methods for addressing the training problem such as using restricted syllable level rescoring on name candidates generated by a phone FSG recognizer. We consider the syllable based recognizer as a first step towards developing a hierarchical recognizer which has units of different size(such as phones, syllables and words). The design methodology of the hybrid syllable recognizer can be extended to multiple level representations by appropriately splitting the lexicon.

Using information retrieval techniques for spoken name recognition allows for a substantial reduction in the recognizer complexity with a small tradeoff in accuracy. As our results indicate the syllable is a very promising unit for such schemes. This can be attributed to the low insertion and deletion rate of syllables. We are investigating techniques which will help us incorporate knowledge of frequent phone addition/deletion/substitution in the statistical search stage. This will help in improving the name candidate list search for the FSG rescoring stage.

We plan to study the performance of our system for names of different linguistic origins, say European, Asian etc. Depending on their linguistic origin names differ widely in their phonetic coverage and average lengths. For e.g. Chinese names are usu-

ally shorter than American names. The effect of these factors on spoken name recognition would be an interesting study.

8. ACKNOWLEDGEMENTS

This research was funded in parts by the Integrated Media Systems Center, an NSF ERC, under cooperative agreement No. EEC-9529152 and by the Department of the Army under contract number DAAD 19-99-D-0046.

9. REFERENCES

[1] R. Billi, F. Canavesio and C. Rullent, "Automation of Telecom Italia Directory Assistance Service:Field Trail Results", *IEEE Workshop on interactive Voice Technology for Telecommunication Applications (IVTTA)*, pp 11-16, Torino, Italy, 29-30 Sept, 1998.

[2] Y.-Q. Gao, B. Ramabhadram, J. Chen, H. Erdogan and M. Picheny, "Innovative Approaches for Large Vocabulary Name recognition", *ICASSP*, pp 333-6, Salt City, Utah, 2001.

[3] Neeraj Deshmukh, Audrey Le, Julie Ngan, Jonathan Hamaker and Joseph Picone, "An Advanced System to Generate Pronunciations of proper Nouns", *ICASSP*, vol. 2, pp. 1467-1470, Munich, Germany, April 1997.

[4] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., and Zavaliagkos, G., "Stochastic pronunciation modeling from hand-labelled phonetic corpora", *Speech Communication*, 2000.

[5] 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.

[6] 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.

[7] S. Greenberg, "The Switchboard Transcription Project", *1996 LVCSR Summer Research Workshop*, Johns Hopkins University, Baltimore, Maryland, USA, August 1996.

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

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

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

[11] 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.

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

[13] 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.

[14] Names 1.1, The CSLU OGI Names Corpus, *http://cslu.cse.ogi.edu/corpora/names/*.

[15] A. Abella, B. Buntschuh, G. DiFabbrizio, C. Kamm, M. Mohri, S. Narayanan, S. Marcus, and R. D. Sharp, "VPQ: A Spoken Language Interface to Large Scale Directory Information", *ICSLP 98* Sydney, Australia, pp. 2863-2867.