ISCA Pronunciation Modeling and Lexicon Adaptation, 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 has traditionally been a difficult task for speech recognition systems because of the large variations in speaking styles, linguistic origins and pronunciation found in names. The 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 thousand names, 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 by syllables can help in improving recognition performance and reducing the system perplexity. We present results comparing systems using 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 30% less than the phoneme-based system. For a phone/syllable level bigram based recognition networks the observed recognition error rate for syllable-based system was about 40% less than the phoneme system.

1. INTRODUCTION

Recognition of spoken names is a critical domain in 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 of the order of several thousands, making the spoken name recognition task a high perplexity 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 tremendous variability in linguitic 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 person of American origin whereas a native speaker from India pronounces it as 'aa b hh ih n eh v'. To increase recognition accuracy a large number of the possible pronunciation variations are typically included in the dictionary.

The problem of pronunciation variability is closely tied with the perplexity problem. Inclusion of multiple pronunciations increases the perplexity substantially, since the recognizer has to now match against multiple paths for every name. Even if computational and memory requirements were not a limiting factor, we would still have the problem of generating a dictionary containing the multiple valid pronunciations of the names. Automatic pronunciation generation techniques based on neural nets or tree based approaches exist to generate the different possible pronunciation 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 a 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 lists. It should be also noted that variations in 'non native' pronunciation also include stress placement or prosody variations which are very difficult to cover in a dictionary.

One basic approach for 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 matching. Thus 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. However it is possible to improve the recognition performance for names which are very common (such as John,Smith) by giving them a higher weight. The perplexity of such recognition networks can be prohibitive for very large name lists comprising 50K or more words. For the spoken name recognition tasks, we observed that as the perplexity of such a recognition network grows the accuracy drops down and for very large namelists the word accuracy becomes comparable to the accuracy of a bigram based phoneme recognition system. One promising approach for cases with large word lists is to use inverse dictionary lookup techniques to get the name. That is we identify the underlying phoneme sequence based on a n-gram (bigram) model and find out the word (name) corresponding to that pronunciation. This can be seen as a Information Retrieval problem. The advantage of such a scheme is that it lowers computational complexity substantially with a small tradeoff in accuracy. Performance of such techniques depends on the accuracy of the n-gram based unit level recognizer and the pronunciation variations covered in the dictionary for name retrieval. For phoneme based systems variations in speaking style and pronunciation degrade the performance of such a scheme substantially.

To address pronunciation and speaker variability in a better way we decided to explore the use of a larger acoustic unit, the syllable. The syllable is a basic unit of speech consisting of two or more phonemes, 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. 2).

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 mixed system which combines syllable and phone based units. In the next section we describe the motivation for using syllables for spoken name recognition and also describe the design of the syllable-based recognizer. Training strategies and corpora are described in section 3. In Section 4 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 with special focus on the research issues we will address in the final paper.

2. SYLLABLE BASED RECOGNIZER

The syllable presents an acoustic unit which because of its extended context information and close relation with human perception of speech, is well suited for dealing with differences in pronunciation, which are common in names. The longer duration of a syllable, which spans multiple phonemes, implies that pronunciations differing in a single phoneme (over a syllabic time span) are less likely to effect the syllabic sequence recognition. In addition to linguistic background differences, significant changes in pronunciation are bought about by phenomena like coarticulation and deletion of phonemes, which are very common in casual speech. The syllable provides a natural framework for integrating coarticulatory phenomena. In studies of the Switchboard corpus [6][7] it has been shown that syllables had a deletion rate of 1% percent whereas the deletion rate for phonemes was 12%. The relative robustness of the syllable against small changes in pronunciation allows us to capture the phoneme level variations in a single syllabic representation. Thus we are able to use single representation for names in a syllable FSG network and keep the perplexity of the network low. In addition, the syllable provides a convenient framework for incorporating suprasegmental prosodic information into the recognition system. A number of psychoacoustic and psycholinguistic studies such as [8][9] have shown that the syllable plays a central role in human speech perception. These factors serve to justify a move from context dependent phoneme-based units to syllabic units.

For very large namelists (100K names or above) where the perplexity of a per-word recognition system would be prohibitive, we can use an Information Retrieval centered scheme for name retrieval using a syllable-based dictionary. The idea is to identify the underlying syllable sequence using a unigram or bigram based syllable recognizer and then do statistical string matching in a syllabic dictionary to identify the right name. We can also use a multipass approach wherein such an N-best list could be used to generate a compact FSG for rescoring. This would outperform similar phone based systems as the dictionary has lesser valid instances (for the same name) to match against (which makes it easier 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 phoneme case.

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 phonemes 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 [10] as a set of rules which define permitted syllable-initial consonant clusters, syllable-final consonant clusters and prohibited onsets. Syllabification software available from NIST [11] implements these rules and comes up with a set of alternative possible syllable clusters given a phoneme 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 phoneme t being shared by two different syllables. However these were very infrequent in the names dataset that we are interested in and we chose to ignore the ambysyllabic information.

The phone level HMM models have the same basic topology with equal number of states, for different phonemes. However syllable models require different number of states depending on their size. A syllable comprising four phonemes such as ‘s_w_eh_l’ requires more number of states than single phonemes or other shorter syllables such as ‘t_eh_n’. To account for this the number of states was chosen to be three per phoneme comprising the syllable. 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 typically name is small.

Another issue with the syllable recognizer is the large number of syllables possible in English, which is hard to cover with even a phonetically balanced speech corpus like TIMIT. Thus when the content of input speech is not well restricted it is important to have a fallback phone level recognizer along with syllable based recognition, so that it can capture speech content for which the syllable-based recognizer has no units. As described in[5], we built a hybrid recognizer in which syllables with only a few occurrences in the training data were replaced by their phoneme level representation in the lexicon,. However performance of this simple formulation is severely limited by two potential factors. First is the lack of context models in the mixed recognizer, which includes phoneme units along with a small number of syllable units. The units corresponding to phonemes are of short duration and without context information their performance is not good. The other factor that leads to relatively poor performance of the mixed recognizer is the acoustic mismatch between syllabic units and phoneme units. Our preliminary results (Sec 4) indeed confirm that this is the case. At present we are investigating other approaches for a mixed recognizer, which use a backoff phoneme recognizer in the case of missing syllabic units.

3. TRAINING: CORPORA AND IMPLEMENTATION

As the first step we built three recognition systems using the TIMIT speech corpus for bootstrapping. For the TIMIT corpus, we had

2800 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 phoneme 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. For the syllable and the mixed system we did not use any context information and all models were single unit. We used 16 Gaussian mixture models for the syllable-based system to allow for its larger acoustic size and contextual variance. After this initial training stage 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[12]. 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 BEEP and 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 triphone based system. For the syllable and mixed system we used the final TIMIT models as the prototypes for the syllables common between the two databases, which are around 1200 in number. The rest were initialized by the standard flat start method. 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 initially trained using TIMIT.

4. PRELIMINARY RESULTS

Considering the broad nature of the spoken name recognition problem, we need to compare the performance of the syllable-based system and the phoneme-based system on different tasks. The first issue is to see how both systems compare on a FSG based spoken name recognition task. Then we compare the performance degradation of both systems when the name list size is increased. For large name lists with 50-100K names, (unweighted) FSG networks incorporating the entire name list are not very practical. To limit the recognizer perplexity in such cases, we propose a two stage approach A bigram based recognition at the phone/syllable level as a first pass followed by statistical string matching in dictionary to form a compact FSG which is then used for recognition. Towards this end, we will compare the unit sequence recognition accuracy of both systems for unigram and bigram based models at the unit level.

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

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

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	Accuracy (%)
Context Independent Phoneme Recognizer	45
Context Independent Mixed Recognizer	58
Context Dependent Phoneme Recognizer	63
Context Independent Syllable Recognizer	75

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

The first task is a complete name list evaluation in which the training vocabulary completely covers the list of names for recognition. This 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. We compared the performance of the three recognition systems on this set. The results are given in Table 4. These results compare well with previous results on similar size name lists[13]. As can be seen from these results for the FSG recognition task, the performance of the syllable-based recognizer is significantly better than the other recognizers. Results comparing the mixed recognizer and context independent phoneme recognizer are presented just for the sake of completeness. As described previously, the design of the hybrid system is still under investigation.

Next we evaluate the recognition performance for large name lists. We will not consider system based on mixed units or context independent phonemes for these evaluations, as their performances are relatively poor. To compare how the performance of the phoneme based and the 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 1 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 phoneme-based system.

For very large word lists we use n-gram based models to identify the underlying unit sequence (phonemes or syllables) and then use reverse dictionary lookup based on statistical string matching to identify the name. We used a bigram model for phonemes as well as syllables and measured the accuracy with which both the systems find out the underlying unit sequence. Syllable based recognizer was able to give a unit recognition rate of 78% whereas the phoneme recognizer accuracy was 62%. 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 phoneme based system.

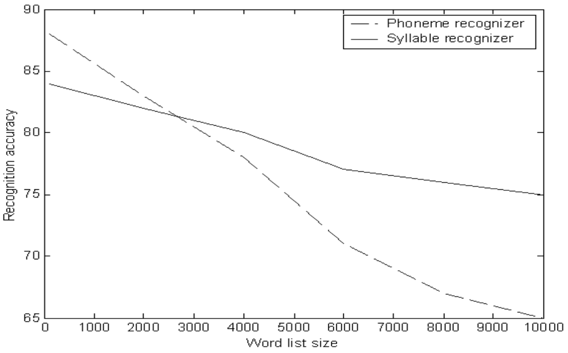

Fig. 1. Plot of recognizer accuracy vs word list size

Our experimental results, which cover a range of spoken name recognition tasks show that a syllable-centered approach outperforms phoneme-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 perplexity.

5. 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 in 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 and speed over context dependent phoneme 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 phoneme 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. However we need to improve on some aspects of our system. We will describe them briefly.

In our current scheme which uses phone/syllable n-gram based language models to address the perplexity issue with large name lists(over 50K), we use reverse dictionary on the best matching sequence to identify the name. A more generic approach is to use N best sequence recognition to get a set of possible unit sequences and then generate a set of candidate names using statistical string matching in the dictionary. The candidate name list can then be used, for example in a FSG framework for spoken name recognition. We plan to report on the effectiveness of such a scheme both in terms of recognition accuracy and speed improvements achieved, in our final paper.

An issue with the syllable recognizer is the large number of syllables possible in English, which cannot be covered even in a large phonetically balanced corpus like TIMIT. In our case we tried to address this by using a mixed recognizer, which uses both syllable, and phone level units. However lack of proper context models affects the performance of the mixed recognizer seriously. We are currently experimenting with other approaches based on N best rescoring to combine phone and syllable units. Results will be presented in the final paper.

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

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

[9] 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.

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

[11] 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.

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

[13] 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.