Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Speech Communication 48 (2006) 1126–1136

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A split lexicon approach for improved recognition of spoken names

Abhinav Sethy [a,*], Shrikanth Narayanan [a], S. Parthasarthy [b]

[a] *Integrated Media Systems Center, Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, CA 90007, United States*

[b] *AT&T Labs-Research, 180 Park Avenue, Florham Park, NJ 07932, USA*

Received 6 December 2004; received in revised form 29 March 2006; accepted 29 March 2006

Abstract

Recognition of spoken names is a challenging task for automatic speech recognition systems because the list of names for applications such as directory assistance tends to be in the order of several hundred thousands. This makes spoken name recognition a very high perplexity task. In this paper we propose the use of syllables as the acoustic unit for spoken name recognition based on reverse lookup schemes and show how syllables can be used to improve recognition performance and reducing the system perplexity. We present system design methodologies to address the problem of acoustic-training data sparsity encountered when using longer length units such as syllables. We illustrate our ideas first on a TIMIT based continuous speech recognition problem and then focus on the application of these ideas to spoken name recognition. Our results on the OGI spoken name corpus indicate that using syllables in place of phoneme models can help boost system accuracy significantly while helping to reduce the system complexity.

Keywords: Syllable; Spoken name recognition; Reverse lookup; Split lexicon

1. Introduction

Spoken name recognition is a key component of many speech recognition applications. Speech based information retrieval relies heavily on accurate spotting of keywords or names. Another common application area is in the call center domain for tasks such as directory assistance, name dialing systems, city name recognition as part of a travel system, caller name identification for banking etc. For most applications, the list of names tends to be on the order of several thousand, making spoken name recognition an inherently high complexity problem. In addition, the large variability in name pronunciation, both at the segmental and suprasegmental levels significantly decreases recognition accuracy. Names have multiple valid pronunciations that evolve as a product of various socio-linguistic phenomena. Specifically in a country like the USA, with a broad cultural base, there is considerable variability in the linguistic

[*] Corresponding author. Tel.: +1 323 3730752.
E-mail addresses: sethy@sipi.usc.edu (A. Sethy), shri@sipi.usc.edu (S. Narayanan), sps@research.att.com (S. Parthasarthy).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.03.005

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'.

Automatic pronunciation generation techniques such as those based on neural nets (Deshmukh et al., 1997), decision trees (Riley et al., 2000), finite state transducers (Galescu and Allen, 2002) and acoustic decoding (Ramabhadran et al., 1999; Beaufays et al., 2003) have been proposed previously. However these techniques require a large set of words and their different pronunciations for training. In addition, performance of such data-driven schemes tends to be limited for names, especially those with foreign 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.

We propose the use of syllable-based reverse lookup schemes to reduce the high complexity of spoken name recognition and show that the larger length of the unit combined with the implicit phonotactic constraints that the syllable model imposes can help boost the accuracy of the reverse lookup scheme substantially. We believe that longer length units such as syllables can help in reducing the dependency on having a representative multiple pronunciation dictionary accurately by modeling larger contexts in acoustics itself. This belief stems from the effectiveness of longer (word) length units in tasks such as digit recognition where word length units remove the need of having a phonetic dictionary with variants. Implicit modeling of pronunciation variations in acoustic models is also supported by Hain (2002). Also these units can help in exploiting longer acoustic correlations beyond the phone. Prosody and stress variations in 'non-native' pronunciation that are very difficult to represent in a dictionary can also be embedded in the acoustic modeling. More details about this line of modeling are given in Section 2.

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 phones, the number of basic models (without context modeling) is fixed for a given language, but if we decide to use syllable or word size units, the number generally increases with the vocabulary size. Many of these units corresponding to infrequently occurring words will have poor coverage in the training data. The sparsity of training data has been the main hindering block in using larger units for tasks such as large vocabulary continuous speech recognition (LVCSR) or spoken name recognition. For small vocabulary tasks such as alphabet or digit recognition, larger units (typically word level units) are used frequently and with success. To address some of the challenges due to data sparsity, in this paper, we present techniques for initialization of larger units using context-dependent (CD) phones which ensure that the system performs well even with minimal or no acoustic data for training the larger units. We also present different criteria to split the lexicon such that we use the appropriate units for representing the vocabulary words with a given training corpus. Proper unit selection is important for reducing system complexity and ensuring robust training.

The paper is organized as follows. Section 2 will discuss syllable-based acoustic modeling. Section 3 describes the general design issues in syllable-based speech recognition. Section 4 describes the design of our spoken name recognition system. The speech corpora used for the experiments in this paper and the training methodology are described in Section 5. Experiments and results are discussed in Section 6. In Section 7, we provide a brief summary of our work, the major findings and an outline for future research.

2. Acoustics based pronunciation modeling using syllables

Differences in speaking styles arising from accent and other factors such as age lead to pronunciation variations which are systematic in nature. Pronunciation variations in spontaneous speech can also occur because of factors such as emotion and coarticulation. It is difficult to represent these effects by pure surface form variants. Longer length acoustic units such as syllables should automatically capture many of these variations occurring at the phone level, thus reducing the dependence on surface form variations. This strategy has worked very effectively for limited domain tasks such as digit recognition where whole word level models are used and hence a dictionary mapping is not required. However, it is not possible to build whole word models for LVCSR systems because of training data sparsity and also because in most cases the test vocabulary is not a subset of the training vocabulary.

Another motivation for using an acoustic unit with a longer duration is that it facilitates exploitation of temporal and spectral variations simultaneously. Parameter trajectory models and multi path HMMs (Gish and Ng, 1996; Korkmazskiy, 1997) are examples of techniques that can exploit the longer duration acoustic context, and yet have had marginal impact on phone-based systems. Units of syllabic duration or longer have the potential to model cross phone spectral and temporal dependencies better than traditional phone-based methods. We are further motivated by results from psychoacoustics research (Greenberg, 1997; Massaro, 1972; Lippmann, 1996) which indicates that syllable length durations play a central role in human perception of speech especially under noise conditions. Recent research on syllable-based recognition (Ganapathiraju et al., 2001; Kirchhoff, 1996; Wu et al., 1988; Shafran and Ostendorf, 2003) also demonstrates that syllables can help in improving performance of ASR.

As explained earlier, the major challenge in using syllables and word level units for recognition is the training data sparsity problem. In (Ganapathiraju et al., 2001) this problem is partially resolved by using only those syllables which have good coverage in the acoustic data. At a context-independent level the syllable, being a larger unit, requires more training data than phone sized units and hence proper training of all syllable models using flat initialization strategies, as described in (Ganapathiraju et al., 2001), is difficult. We describe a simple scheme for initializing syllable models from context-dependent (CD) phones which helps in robust estimation of syllable models. In addition, we need to estimate the advantage, in terms of recognition accuracy, that we can gain by moving from a phone representation to a syllable or a whole word representation. In general the achievable improvement depends heavily on the training data, since it determines how well longer acoustic units can be trained. Thus depending on the acoustic-training data available to us we need to determine the proper representation for each word in the lexicon. We describe two methods of creating a lexicon representation for variable-length acoustic units. Section 3 will describe the proposed longer-unit initialization and lexicon splitting strategies in more detail.

3. Recognizer design

In the first stage, we design three separate recognizers corresponding to the different acoustic units of interest i.e. phone, syllable and word.

3.1. Phone recognizer

The design of the phone-based recognizer follows the standard flat start (i.e. uniform segmentation) Baum–Welch reestimation strategy with decision tree based within-word triphone creation and clustering (Odell et al., 1995). The phone-based recognizer serves as a baseline for our results and is also the only context-dependent system that we built. Instead of flat start initialization we can use prebuilt models from an existing corpus to initialize the segmentations for the corpus of interest.

3.2. Syllable recognizer

The first stage in designing a syllable lexicon is to syllabify the training vocabulary and generate the list of syllables present in the training data. To do this we need to convert the phone level pronunciation in the dictionary to the corresponding syllabic representation. This process of syllable boundary detection (syllabification) is described in (Kahn, 1976) as a set of rules which define permitted syllable-initial consonant clusters, syllable-final consonant clusters and prohibited onsets. Syllabification software available from NIST (Fisher et al., 2000) implements these rules and given a phone sequence comes up with a set of alternative possible syllables which can be used to generate the syllabic pronunciation. In our system, 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 phones comprising the syllable. For example *Junior* with the phonetic transcription ‘jh uw n y er’ can be represented in syllabic terms as ‘jh_uw_n’ ‘y_er’. Homophones were given the same lexical representation.

It is possible for a given phone sequence to have multiple valid syllabifications due to consonants at syllable boundaries that belong, in part, to both the preceding and the following syllable. These consonants are referred to as ambisyllabic. In our case, for simplicity, we chose the most commonly occurring syllabification, ignoring ambisyllabicity. Using multiple syllabifications to represent a single word in the dictionary might potentially decrease the training data per unit substantially. It is also possible that other syllabifications provide better fit to the speaking style and rate of particular speakers. The effect of ambisyllabic representations on recognition performance is a research issue, which was not addressed in this paper.

Phone level HMM models typically share the same topology with the same number of states for all phones. However, syllable models require a different number of states depending on their size. A syllable comprising four phones such as 's_w_eh_l' requires more 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.

To initialize the models for the syllable recognizer we used context-dependent (CD) phone models trained on the same data. Context-dependent phone states were concatenated to generate the states of the corresponding syllable models (Fig. 1). The context for each phone model was restricted to the surrounding phones inside the syllable. The same principle can be extended to generate word level models. It should be noted that in our design the syllable models are copied independent of the words in which they occur and we did not use context information beyond the syllable boundary. For example, even though the left context of 'ey' for the syllable 'b_l_iy' in 'ably'(ey b l iy) would be a more appropriate choice while copying the model for the phone b (ey-b+l instead of b+l), we restrict ourselves to the context-independent model (b+l) only since the same syllable might occur in some other word with a different context.

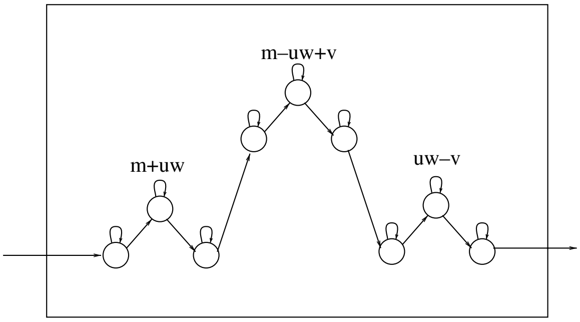

Fig. 1. Initialization of the nine state syllable m_uw_v.

3.3. Larger acoustic units

Longer length units such as whole word models can be built in the same fashion as the syllable models. For monosyllabic words the syllable and word models are equivalent. We did not build separate word level models for monosyllabic words as suggested in (Ganapathiraju et al., 2001).

Initialization from phone level models in this manner ensures that the syllable and word level models perform similar to the corresponding phone recognizer even without further acoustic training. The increase in recognition accuracy that can be achieved by preferring a syllabic unit to the phone representation depends on the coverage of the unit in the training data. This brings out the need to identify the proper lexical representation or choice of units to represent the words in the lexicon. We tried two different strategies for addressing this problem. The first uses a simple threshold based on the number of training units available in the acoustic data. The second uses the difference in recognition accuracy achieved on the training data by using syllable, phone or word level units. Based on this unit selection process, we split the word lexical entries appropriately. Certain commonly occurring words such as 'the' will automatically end up having word level models. Other words will have either a pure syllable representation or a hybrid syllable and phone representation. As an illustrative example, consider the phone level representation 'ae k t x r z' for 'Actors'. The pure syllabic representation would be ae_k t_x_r_z. However if the syllable t_x_r_z is not included in the list of acoustic units based on the selection criteria mentioned above, we split the lexicon entry as 'ae_k t+x t-x+r x-r+z r-z' with '-' referring to left context and '+' to the right context. Our results indicate that the second approach of using training data recognition accuracy as a unit selection criterion achieves better performance than the counts based scheme.

4. Spoken name recognition systems

The intended application for the aforementioned modeling in the paper is spoken name recognition. In this section we will describe our decoding strategies with reference to the spoken names recognition task. 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 that 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 where all entries are deemed equally likely (Abella et al., 1998; Billi et al., 1988). Considering this we gave equal weight to all the names although it is possible to improve the recognition performance by giving a higher weight to common English names such as James, Mary, John and Patricia. 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 100 K or more words for many directory applications.

For spoken name recognition tasks, we observed that as the perplexity of the names FSG recognition network grows, the phone level recognition accuracy drops along with the WER. The phone recognition accuracy achieved with the lexical constraints imposed by the FSG network comes close to the accuracy of a simple phonotactic model based recognizer. Based on this observation, a promising approach (Fig. 2) 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 an n-gram (phone) language model and then use statistical string matching to find 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 pronunciations. Knowledge of frequent phone insertion, deletion and substitution can be incorporated in the statistical search stage to make it more accurate. The current implementation however is limited to taking the Levenshtein (i.e. string edit) distance.

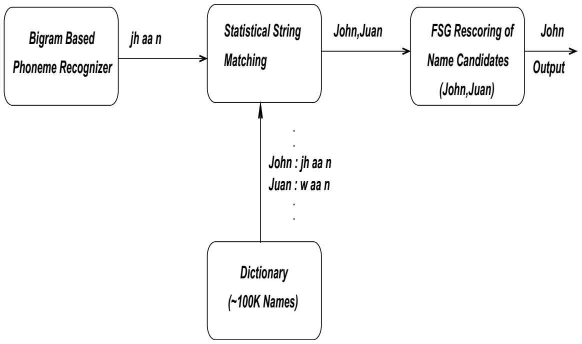

Fig. 2. Information retrieval scheme for name recognition provides scalability.

The N-best list of name candidates (or equivalent lattice) generated by the inverse lookup stage 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 trade-off 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. The phone is not a good unit for information retrieval based name recognition schemes because of factors such as non-native pronunciation variations and the high rate of phone insertion 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. The syllable is a better unit for reverse lookup based schemes. The syllable sequence recognition accuracy is higher than the phone sequence recognition accuracy because the syllable constrains the decoding process. The reverse lookup process becomes equivalent to the FSG decoder if we use word length units. The syllable provides a compromise between the fast but very inaccurate phone sequence recognizer and the prohibitively slow word sequence recognizer (FSG network), which provides the upper bound on the accuracy possible in this decoding scenario. An additional advantage of the reverse lookup scheme is that by reducing the name candidate list it provides the flexibility of using more complex algorithms, which would have been computationally prohibitive for a large FSG network. The reverse lookup approach is similar to the retrieval schemes used in spoken document retrieval (Thong et al., 2000), OOV and named entity detection (Geutner et al., 1988). However by decoding with syllables we implicitly add syllable length phonotactic constraints which coupled with the larger length of the unit (covering more acoustic context) helps in boosting the recognition accuracy.

We will next describe the spoken names corpus that we used in our experiments and the training setup.

5. Training for spoken name recognition: corpora and implementation

5.1. Initial TIMIT system

As the first step we built three independent recognition systems (phone, syllable and hybrid) using the TIMIT speech corpus. For the TIMIT corpus, we had 3137 syllables with about 70% of the words being either monosyllabic or bisyllabic (Table 1).

The speech data from TIMIT was down-sampled to 8 kHz. Twenty-six mel frequency cepstral coefficients were extracted at a frame rate of 10 ms using a 16 ms 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 from the CMU phoneset were initialized and trained on hand labelled 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 eight Gaussians per state. Subsequently the syllable and the hybrid system were initialized as described in Section 3 and were trained on the acoustic data. We used the TIMIT models to seed the models for the NAMES corpus as described in Section 5.2.

5.2. NAMES corpus

The primary speech corpus of interest to us for spoken name recognition is the OGI NAMES corpus (Sethy et al., 2002). 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 were not reading the names off a list. Word level transcriptions are provided for all name utterances and some of the utterances are also labelled phonetically. The phonetically labelled files were used to make a names dictionary, which was augmented with some additional name entries from public domain dictionaries like Cambridge University's BEEP dictionary and the CMU dictionary. The names corpus is sampled at 8 kHz and has about 6.3 h of speech data. There are about 10,000 unique names in the corpus with a rich phonetic coverage which amounts to 40% of the possible phone pairs. Tables 1 and 2 describe the occurrence frequency for words of different syllabic count in the TIMIT and NAMES corpora. As can be seen, most names are bi or tri syllabic unlike in TIMIT, which has a higher monosyllabic content mainly due to functional words such as 'and' and 'the'. Also, words with smaller syllable count are used more frequently in generic sentences of the nature found in TIMIT. Syllable count distributions for a conversational speech corpora such as Switchboard can be seen in (Ganapathiraju et al., 2001).

Table 1
Distribution of words and their syllable count for the training part of the TIMIT corpus

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

Total number of words in this training corpus was 40,000.

Table 2
Distribution of words and their syllable count for the NAMES corpus

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

Total number of words was 10,000.

5.3. Bootstrapping from TIMIT

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 to build a NAMES CD phone system. For the syllable and the hybrid unit system we used the final TIMIT models as prototypes for the syllables common between the two databases, which are around 1200 in number. Table 3 shows the distribution of

Table 3
Distribution of syllables common to TIMIT and NAMES and their length

	Syllable length			
	2	3	4	5
Percentage of common syllables	30	53	15	2

Total number of common syllables is around 1200. This table does not include single phone syllables.

the syllables common between TIMIT and the NAMES database for various syllable lengths. We can see that a large number of the shorter syllables (2–3 phones in length) can be initialized from TIMIT. Single phone syllables such as "ax" (as in "about" ax b_aw_t) are not shown in Table 3 as they can be initialized directly from phone models. The remaining syllables in the NAMES lexicon were initialized using the techniques described in Section 3 from the NAMES CD phone models.

6. Results and discussion

After training, the first set of recognition experiments was conducted on the TIMIT test set to verify our ideas. Subsequently we focused on evaluation of our system on the NAMES recognition task.

6.1. Preliminary TIMIT based experiments

We developed a hybrid unit speech recognition system, which was trained and evaluated on the TIMIT corpus using the train and test sets provided. The TIMIT corpus contains a total of 6300 sentences with 1344 sentences in the test subset. A basic word bigram language model trained on TIMIT was used in our experiments. We built two sets of models, the first one used CD phone initialization of syllables as described in Section 5. For performance comparison purposes we also built syllable and word level systems using the standard flat start and embedded training of the acoustic models (Ganapathiraju et al., 2001).

6.1.1. Performance improvements through larger acoustic units

We compared the performance of recognizers using the syllable and word level units with the phone-based system. Using CD phone system based initialization guarantees that the syllable and word level systems perform similar to the baseline phone case even without reestimation. In Table 4 we show the recognition accuracy obtained with three sets of acoustic models—context-independent syllable, context-independent word and context-dependent phone models at different stages of parameter reestimation. The initial accuracy is identical to the CD phone system for the word case and is slightly lower for syllable models which can be attributed to the lack of context modeling across syllable boundaries. The recognition accuracy subsequently improves with reestimation for units which have significant coverage in the training data. The context-dependent phone models were not reestimated as no performance gains could be observed by additional rounds of reestimation.

We compared the word recognition accuracy achieved with syllable and word systems trained using the standard flat start strategy. As can be seen in Table 5, the choice for the initialization strategy makes a significant difference in performance. Assuming that a typical syllable or word level model will have three or more phones, the number of parameters to be estimated for the model is around three times that for the phone models. Thus a large number of units in the flat start method are poorly trained. An analysis of the recognition errors confirms that the performance difference between the flat start method and CD phone-based initialization can be attributed to units which do not occur frequently in the training data.

6.1.2. Hybrid lexical unit recognizer

An analysis of recognition performance on the training data was used to decide which representation of a word is best. In our present implementation we choose from pure syllabic, pure word or pure phone representations based on what representation gave a better WER during the training data decoding. We also tried an alternative simpler scheme

Table 4
TIMIT word recognition accuracy results with syllable and word level units at different stages of reestimation after CD phone initialization compared to baseline phone recognizer

Recognizer type	First reestimation (%)	Third reestimation (%)
Context-independent syllable	80	85
Context-independent word	82	87
Context-dependent phone	82	82

Table 5
TIMIT word recognition accuracy results for syllable and word level units with and without CD phone-based initialization after three reestimations

Recognizer type	Flat initialization (%)	CD phone-based initialization (%)
Context-independent syllable	80	85
Context-independent word	81	87

Table 6
Word recognition accuracy for TIMIT and complexity in number of states of syllable, word and hybrid lexicon recognizers

Recognizer type	Accuracy (%)	Number of model states in recognizers
Context-independent word	87	43,380
Context-independent syllable	85	24,460
Hybrid unit recognizer	90	13,450

which uses instance counts to do the model selection. The word representation was preferred over the syllable representation if the word occurred more than a hundred times in the training data and the syllable was preferred over the phone if the occurrence count was more than 75. The performance of the counts based scheme was found to be lower than the first scheme and involved experimentation with the count cut-off thresholds. Thus we discarded this approach and used the first scheme in our experiments.

The complexity of the recognizers was evaluated in terms of the total number of states the models required (see Table 6). The total number of states is linked to the memory requirements and speed of the Viterbi decoder. As expected, the hybrid unit recognizer has a substantially lower complexity as compared to the syllable or word recognizers. Interestingly the hybrid unit recognizers also had a slightly higher accuracy. We can attribute this to the unit selection process, which ensures that the hybrid recognizers include only those units which are robustly trained. It is important to note that the complexity in terms of the physical model states is much lower for the context-dependent phone recognizer. The CD phone recognizer had around 4000 distinct physical states and about 18,000 logical states. No form of state tying was used for word and syllable models. It should be noted that for linear decoding systems such as HTK in FSG mode, which do not minimize the decoding graph at the acoustic unit level, the decoding complexity is independent of the number of states and depends just on the number of words. For these recognizers the number of states should be taken as a measure of the acoustic model memory requirements only.

The TIMIT corpus is designed primarily for phone recognition experiments and the results should be interpreted with the proverbial pinch of salt. The performance improvements on TIMIT that we could achieve using syllables over the context-dependent phonetic system are much higher than what we could achieve in LVCSR tasks. Consider, for example an LVCSR system where syllable-based models were investigated more recently. For the MALACH automatic speech transcription system designed for spontaneous interview speech from holocaust survivors (Sethy et al., 2003) the WER improvement achieved by using syllables was only about 2% (in relative terms) when compared to the corresponding phonetic system. However we were still able to achieve a 15% improvement in recognizing keywords and names, a task which is similar in nature to the spoken name recognition task we have addressed here.

6.2. Evaluation on the spoken name recognition task

Comparative performance evaluation between the syllable-based and phone-based systems is presented for both the FSG network based spoken name recognition and the information retrieval schemes. As discussed in Section 3, 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 referred 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 referred to as the hybrid recognizer. The initialization and training of all models were done in the manner described in Section 5 using both the TIMIT and NAMES databases.

6.2.1. Scheme I: FSG networks

As a first step, we performed comparative evaluation of the phone-based recognizer with the syllable and the hybrid recognizer for a FSG based recognition task. In the case of the hybrid recognizer, the syllable was preferred over the phone if the occurrence count was more than 50. The training vocabulary completely covered the list of names for recognition. We randomly selected a section of the NAMES database comprising 6000 utterances 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 7. The results for phone models compare well with previous results on similar size name lists (Abella et al., 1998).

As can be seen from these results for the FSG recognition task, the performance of the two sylla-

Table 7
Recognition rates for different FSG based spoken name recognition systems on the 6000 utterance test set

Recognizer type	Word recognition accuracy (%)
Context-independent phone recognizer	45
Context-dependent phone recognizer	72
Context-independent hybrid recognizer	75
Context-independent syllable recognizer	77

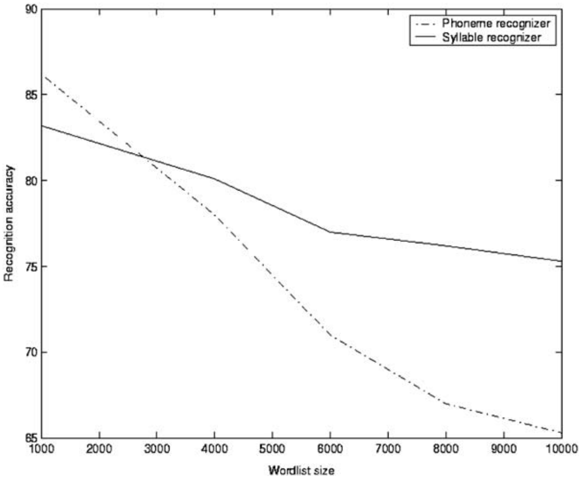

Fig. 3. Plot of recognizer accuracy vs word list size for phoneme and syllable recognizers.

ble-based recognizers is significantly better than that of 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 1 K names and increased the size of the test name list from 1 K to 10 K. Fig. 3 shows the recognition accuracy with increasing vocabulary size. As can be seen from the figure, the (rate of) performance drop for the syllable-based system is less than the drop for the phone-based system.

6.2.2. Scheme II: Information retrieval

In the case of very large name lists, we first used phone and syllable N-gram decoding graphs to identify the underlying unit sequence (phones or syllables) and then used reverse dictionary lookup based on statistical string matching to identify the name (see Section 4). In the first stage for reverse lookup, we identified the two best unit sequence candidates. We used bigram phone and syllable language models estimated on the NAMES corpus to do phone and syllable sequence recognition. In the next stage a name candidate list was selected using reverse lookup in the dictionary (Section 4). For the syllable recognizer we split the recognizer output sequence to the corresponding phone sequence in order to calculate the Levenshtein distance. All names whose distance was less than a threshold were selected for creating the FSG network for the second stage. The FSG network assigns equal weight to all the candidate names. The threshold for selection was taken to be a function of the recognized sequence length and the recognizer type. The recognition speed of reverse lookup scheme depends on the computational complexity of the first stage and the number of name candidates generated for the second stage FSG rescoring. The simplicity of the phonotactic decoding network leads to a very fast operation for the first stage. However the first stage output has low accuracy and consequently a large number of candidate names need to be considered in the second stage FSG rescoring to maintain accuracy. The syllable system has relatively high complexity for the first stage but the second stage FSG size can be much smaller because of the higher accuracy of syllable sequence recognition (Section 4). We choose the distance threshold such that the two recognizers had about the same decoding speed. As can be seen from our results in Table 8 the syllable recognizer has a significantly higher accuracy at the same decoding speed.

6.3. Syllables and pronunciation variation in names

The OGI corpus is woefully lacking in terms of non-native content, making it difficult to conclusively state that syllables help in modeling the pro-

Table 8
Spoken name recognition accuracy for the information retrieval scheme after FSG rescoring of compacted name list

Acoustic unit	Word recognition accuracy after FSG rescoring (%)
Context-independent phone	45
Context-dependent phone	61
Context-independent syllable	73

nunciation variation found in non-native names. To study this further, we are in the process of collecting a spoken names corpus at USC which will have a large chunk of names of foreign origin spoken by both native and non-native speakers of English drawn from the student population at USC. The observation that syllables help in modeling pronunciation variation for names is supported by our results on the MALACH corpus (Sethy et al., 2003) which has a large collection of foreign names. Compared to the overall 2% WER improvement, the WER improvement for names was about 15% with the syllable-based models over context-dependent phones.

7. Conclusion

In this paper we presented a case for using syllables for spoken name recognition. We also implemented a system trained and tested on TIMIT, which uses syllable and word level units in conjunction with CD phone units. To address the problem of training data sparsity for word and syllable units, we used CD phone-based initialization which guarantees that the higher level units will give equivalent performance to CD phone recognizer even in the absence of any acoustic-training data.

Our results which compare the performance of CD phone-based system with syllable and word level systems for the TIMIT and the NAMES recognition task show that substantial performance gains can be achieved by using longer length acoustic units. However given the limited nature of the vocabulary and language models used, the TIMIT experiments we conducted are not really indicative of a full scale LVCSR system. We implemented a similar approach to syllable design in (Sethy et al., 2003) for the MALACH task and found out that in general WER gains from using syllables were much less pronounced than the gains observed specifically in terms of spotting keywords and names.

Using information retrieval techniques for spoken name recognition allows for a substantial reduction in the recognizer complexity with a small trade-off 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 insertion/deletion/substitution in the statistical search stage. This will help in improving the name candidate list search for the FSG rescoring stage.

We believe that using larger units can also help in solving the spoken name pronunciation generation. Names have varied pronunciations and in tasks such as directory assistance the name lists may have more than 100 K names, making it impossible to have manual generation or verification of the pronunciation dictionaries. Extending the current phone-based techniques for pronunciation generation such that they use variable length units (demisyllables/syllables/words) would open the possibility of compensating phonemic representation ambiguity in the acoustic model itself.

We are in the process of collecting a spoken names corpus containing names from different linguistic origins spoken by native and non-native speakers of English. We believe that this corpus will help in better comparison of the performance of syllables with phonetic models in dealing with the pronunciation variation found in non-native speech. 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 for the future.

Acknowledgments

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

References

Abella, A., Buntschuh, B., DiFabbrizio, G., Kamm, C., Mohri, M., Narayanan, S., Marcus, S., Sharp, R.D., 1998. VPQ: A spoken language interface to large scale directory information. In: Proc. ICSLP.

Beaufays, F., Sankar, A., Williams, S., Weintraub, M., 2003. Learning linguistically valid pronunciations from acoustic data. In: Proc. Eurospeech.

Billi, R., Canavesio, F., Rullent, C., 1988. Automation of telecom italia directory assistance service:field trail results. In: IEEE Workshop on Interactive Voice Technology for Telecommunication Applications (IVTTA).

Deshmukh, N., Ngan, J., Hamaker, J., Picone, J., 1997. An advanced system to generate pronunciations of proper nouns. In: Proc. ICASSP.

Fisher, M., 2000. Syllabification Software, The Spoken Natural Language Processing Group, National Institute of Standards and Technology, Gaithersburg, MD, USA [Online]. Available from: <http://www.nist.gov/speech/tools/tsylb2-11tarZ.htm>.

Galescu, L., Allen, J., 2002. Pronunciation of proper names with a joint *n*-gram model for bi-directional grapheme to phoneme conversion. In: Proc. ICSLP.

Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G., Picone, J., 2001. Syllable-based large vocabulary continuous speech recognition. IEEE Trans. Speech Audio Process. (May).

Geutner, P., Finke, M., Waibel, A., 1988. Phonetic-distance-based hypothesis driven lexical adaptation for transcribing multlingual broadcast news. In: Proc. ICSLP.

Gish, H., Ng, K., 1996. Parameter trajectory models for speech recognition. In: Proc. ICSLP.

Greenberg, S., 1997. On the origins of speech intelligibility in the real world. In: Proc. ESCA Workshop on Robust Speech Recognition for Unknown Channels, April.

Hain, T., 2002. Implicit pronunciation modelling in asr. In: Proc. PMLA.

Kahn, D., 1976. Syllable-based Generalizations in English Phonology. Ph.D. dissertation, Indiana University Linguistics Club, Bloomington, IN, USA.

Kirchhoff, K., 1996. Syllable-level desynchronisation of phonetic features for speech recognition. In: Proc. ICSLP.

Korkmazskiy, F., 1997. Generalized mixture of HMM's for continuous speech recognition. In: Proc. ICASSP.

Lippmann, R., 1996. Speech perception by humans and machines. In: Proc. Workshop on the Auditory Basis of Speech Perception, July.

Massaro, D., 1972. Perceptual images, processing time and perceptual units in auditory perception. Psychol. Rev. 79, 124–145.

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

Ramabhadran, B., Deligne, S., Ittycheriah, A., 1999. Acoustics-based baseform generation with pronunciation and/or phonotactic models. In: Proc. Eurospeech.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., 2000. Stochastic pronunciation modeling from hand-labelled phonetic corpora. Speech Comm. 29, 209–224.

Sethy, A., Narayanan, S., Parthasarthy, S., 2002. A syllable based approach for improved recognition of spoken names. In: Proc. ISCA Pronunciation Modeling Workshop.

Sethy, A., Narayanan, S., Ramabhadran, B., 2003. Improvements in English ASR for the MALACH project using syllable-centric models. In: Proc. IEEE ASRU.

Shafran, I., Ostendorf, M., 2003. Acoustic model clustering based on syllable structure. Comput. Speech Lang. 17 (4), 311–328.

Thong, J.V., Goddeau, D., Litvinova, A., Logan, B., Moreno, P., Swain, M., 2000. Speechbot: a speech recognition based audio indexing system for the web.

Wu, S.-L., Kingsbury, B., Morgan, N., Greenberg, S., 1988. Incorporating information from syllable-length time scales into automatic speech recognition. In: Proc. ICASSP.