PHONOLOGICAL RULES FOR ENHANCING ACOUSTIC ENROLLMENT OF UNKNOWN WORDS

Bhuvana Ramabhadran and Abraham Ittycheriah

IBM T. J. Watson Research Center, P. O. Box 218, Yorktown Heights, NY.
email:bhuvana@watson.ibm.com, phone: (914)-945-2976

ABSTRACT

Phonetic baseforms are the basic recognition units in most speech recognition systems. These baseforms are usually determined by linguists once a vocabulary is chosen and not modified thereafter. However, several applications, such as name dialing, require the user be able to add new words to the vocabulary. These new words are often names, or task-specific jargon, that have user-specific pronunciations. This paper describes a fast method for generating clean, phonetic transcriptions (baseforms) of words using phonological rules on baseforms that have been derived based on acoustic evidence alone [8]. It does not require any prior acoustic representation of the new word, is vocabulary independent, and uses phonological rules in a post processing stage to enhance the quality of the baseforms thus produced. Our experiments demonstrate the high decoding accuracies obtained when baseforms deduced using this approach are incorporated into our speech recognizer. Our experiments also compare the use of acoustic models that are trained on task-specific data with models trained for a general purpose (digit, names and large vocabulary recognition) for the purpose of generating phonetic transcriptions.

1. INTRODUCTION

There has been considerable interest in telecommunications based speech recognition services that provide user configurable vocabularies. Name dialing is one such example of a telephony application, where it is necessary to have the ability to provide speaker dependent vocabularies for repertory dialing. This feature will enable the user to add a word(s) to their personalized vocabulary, for which an a priori spelling or acoustic representation does not exist in the speech recognition system, and associate that word(s) to a phone number to be dialed. Once the personalized vocabulary is configured, the user can subsequently dial the phone number by speaking the new word(s) just added to the vocabulary. In order to do this, proper deduction of the phonetic baseform is essential. An automated method for generating speaker dependent acoustic representations (baseforms) from speech in terms of speaker independent subword acoustic units (phones), in order to build the personalized vocabulary was presented in an earlier paper [8]. Once an initial set of baseforms are obtained in a fashion described in [8], phonological rules are applied to this set to obtain additional baseforms from one initial enrollment utterance. The primary use of these rules are to eliminate illegal phone transitions that appear in the final phonetic stream due to noisy environments and the elimination of fricatives and nasals that are introduced in the utterance, within and between words. These refinements include rules to cover tensing, offglide adjustments, presence of fricatives, glottalization, aspiration of stops, merger of vowels before nasals, unstressed vowel reduction, glottal stop insertion, deletions in consonant clusters, desyllabification of high vowels, etc. The phonological rules were derived after a careful analysis of the most frequent errors made by the recognizer. These linguistic constraints, particularly helped in obtaining cleaner baseforms for short utterances, especially when they were uttered from a very noisy background, such as in an airport, work floor or a cafetaria.

Using a sample utterance of the word to be added, a phonetic representation is extracted using the technique described above. This corresponds to one pronunciation for the word enrolled by the user that will be used subsequently for recognition. Subsequent utterances of this word are decoded by the speech recognition system using the newly added representations of the word.

Many speech recognition systems (desktop applications, such as dictation transcription, and telephony applications, such as name dialing) provide the user with the ability to configure personalized vocabularies. Existing methods have shortcomings. The most typical method cannot be used for telephony applications as it requires the spelling input from the user. Furthermore for many words, the spelling of a word has very little correlation with its pronunciation. There are many telephony implementations in service over the public telephone network today [7].

The technique described in this paper for generating phonetic transcriptions (baseforms) for obtaining speaker dependent word pronunciations falls under the second category. For the name-dialing task, we obtain good recognition accuracies during the retrieval phase with only one utterance used during the enrollment procedure. The baseforms thus generated is used by the speech recognizer during recognition, when the personalized vocabulary is active along with other speaker independent navigational commands such as "forward my calls to", "delete name", etc. The accuracy of the recognizer further improves if the bigram statistics used to train the ballistic labeler are obtained from a names corpora as opposed to general purpose database. There is an additional improvement if an additional utterance is used during the enrollment procedure.

The structure of this paper is as follows. Section 2 describes the ballistic labeler algorithm used for the generation of phonetic baseforms followed by the phonological rules used to obtain additional, cleaner baseforms. Section 3 describes a set of experiments that evaluate the recognition performance of the speech recognizer that uses the automatically deduced baseforms. Section 4 discusses the results and some ongoing work and has suggestions for future work.

2. ALGORITHM FOR GENERATING BASEFORMS

This section describes the algorithm that is used for generating the phonetic baseforms from the acoustics alone. The goal here is to find the phone string P that maximizes $p(P|U)$, where

U is the utterance for which the baseform is to be generated. The algorithm proceeds as follows. The acoustic data from the enrolled utterance is labeled in less than real time using the ballistic labeler. This involves the construction of a trellis of arc (sub phone units) nodes from the speech utterance. The probability of a transition occurring from one arc to another is determined by weighting the score obtained from a Hidden Markov Model (HMM) [3] with a precomputed arc to arc transition probability obtained from any training corpora. We show results that compare the use of a general large-vocabulary training corpora and a task-specific, names-only training corpus for this purpose. At any time frame the set of active nodes in the trellis is defined as the nodes with scores greater than a certain pruning threshold. Once the entire utterance has been processed, a back-tracking procedure is employed that traces the best arc-predecessor from the end of the utterance, forcing silence at the beginning and the end of the utterance. Thus, a sequence of phonetic arcs are obtained from which a phone sequence (baseform) is derived for that enrolled utterance. This corresponds to one pronunciation for the word enrolled by the user that will be used subsequently for recognition [6]. To this pronunciation, a set of phonological rules (described in Section) are applied and the resulting set of baseforms are added to the first one.

2.1. The Ballistic Labeler

The ballistic labeler is used to derive the phonetic strings from the acoustic vectors and the precomputed arc transition probabilities [8]. This process can be done in real time when performed at the arc (a phone is made up of 3 sub phone units we refer to as arcs which correspond to the states of the HMM), though for better accuracy its preferred to do it at the leaf level. The leaf which is a context dependent arc is an instance of the arc trained by using the context in which a vocabulary word appears. For example, in the system described below, there are 156 arcs and 2448 leaves.

An example of an arc sequence and a phonetic sequence derived thereafter for a sample utterance 'MOM' is illustrated in Figure 1.

2.2. Phonological Rules

To the baseform sequence obtained above, phonological rules are applied as a post-processing phase. This helps in capturing speech from noisy segments and also in modeling coarticulation that exists in normal speech. These rules were arrived at after analyzing the errors made in a development test set during the recognition phase. For example, when a speaker enrolled 'DAVID CHADDY', the baseform 'D$ T EY B IX DD T AH DX IX Z D$' was generated. During recognition, this was confused with another 'DAVID TOPEY' which had a baseform 'D$ D EY DX IX DD T OW P IY F S D$' that existed in this speaker's personalized vocabulary. Both these names are highly confusible and had been enrolled in a noisy environment which led to the formulation of a rule to capture noisy segments. A second example is the baseform for the utterance 'DAD'. The baseform obtained was 'D$ D EH D T S V F D$'. Again, the call was made from a cellular phone during enrollment, therefore, even though a correct baseform for 'DAD' was obtained, the string of consonants that were generated to cover the noisy frames of speech, force the recognizer to always look for a match in subsequent calls for a lengthier utterance. Since this is not true in a clean environment, we will now confuse 'DAD' with any other short word in the person's vocabulary. In an attempt to reduce errors like this and other coarticulation effects, phonological rules were used, a sample set of which is given in Table 2.

Method Employed	DB I	DB II
Baseforms from 1 acoustic utterance	96.4%	98%
Baseforms from 2 acoustic utterances	96.6%	98.2%
Hand-written Baseforms	97%	94.5%

Table 1: Recognition performance for two databases using baseforms generated with acoustics alone and hand-written baseforms

3. EXPERIMENTS

This method of generating baseforms was evaluated on the name-dialing task, using an in-house data collection software for telephony data collection. Two local databases were built for evaluating the baseform generation algorithm. The first database(DB I) was built using ten speakers and each speaker asked to enroll twelve different names. These names were chosen at random from a currently operational name-dialing application and were complicated enough to produce user-specific pronunciations (for example, CELESTINO DOMINGUEZ, THOMAS CABAN, TONY VAIANISI, etc.). Each participant made ten calls, from as many different phones as possible. These included cellular, digital and analog phones under different environment conditions, such as hallways and the cafeteria. The vocabulary for the recognition task included these names along with other navigational commands such as 'forward my calls', 'call return', 'delete a name', etc. resulting in a 65 word vocabulary. For the second database (DB II), ten speakers with a variety of accents called in from three different phones (digital, cellular and speaker phone conditions). These speakers were asked to read fifty items, that included names, such as, ROSE, ANTHONY FABRIZIO, DAD, MOM, etc.), and other command words, such as, 'HELP', 'CANCEL', 'CALL RETURN', etc.. The vocabulary for this database was made up of 105 words. For both the databases, baseforms were generated from one speech utterance and the remaining calls (9 for DB I) were used as the test bed for recognition with the newly generated baseforms. The decoding results obtained when baseforms were generated with one acoustic token (utterance) and two tokens are tabulated in Table 1. In this experiment, the acoustic models were trained on a names-only training corpus. The numbers presented in the table are the average percentages obtained over all the speakers under all the conditions under which the data were collected. While it can be seen that the recognition accuracy improves marginally with the use of an additional enrollment utterance, the accuracy of the recognizer is very good for the newly added words even when one utterance was used for the baseform generation.

In the above experiment, acoustic models were built from a domain-specific database, i.e. 400 speakers recording 12000 names. These models gave a relative high-degree of accuracy in deriving the phonetic representations. Since the HMMs are trained at the context-dependent feneme level, the state transitions are strictly constrained by the nature of the data. There are a lot of weird transitions present in names, which are not seen in a more general purpose large-vocabulary speech recognition system. This can be seen in the sparseness of the transition matrix obtained using the two models. In a situation, where it is necessary to include various domains in the recognition task, such as, digits, commands, dictation, general yellow-pages search, etc., it is preferred to use one set of models for all these domains. Since the quality of the baseforms obtained using these models is inferior compared to a set of HMMs trained on name-specific data, they can be refined further by using phonological rules. Phonological rules have also proven to be very useful in modeling coarticulation for spontaneous speech recognition. Not only are the baseforms thus obtained cleaner (refer example shown below), they also help in better recognition using these models.

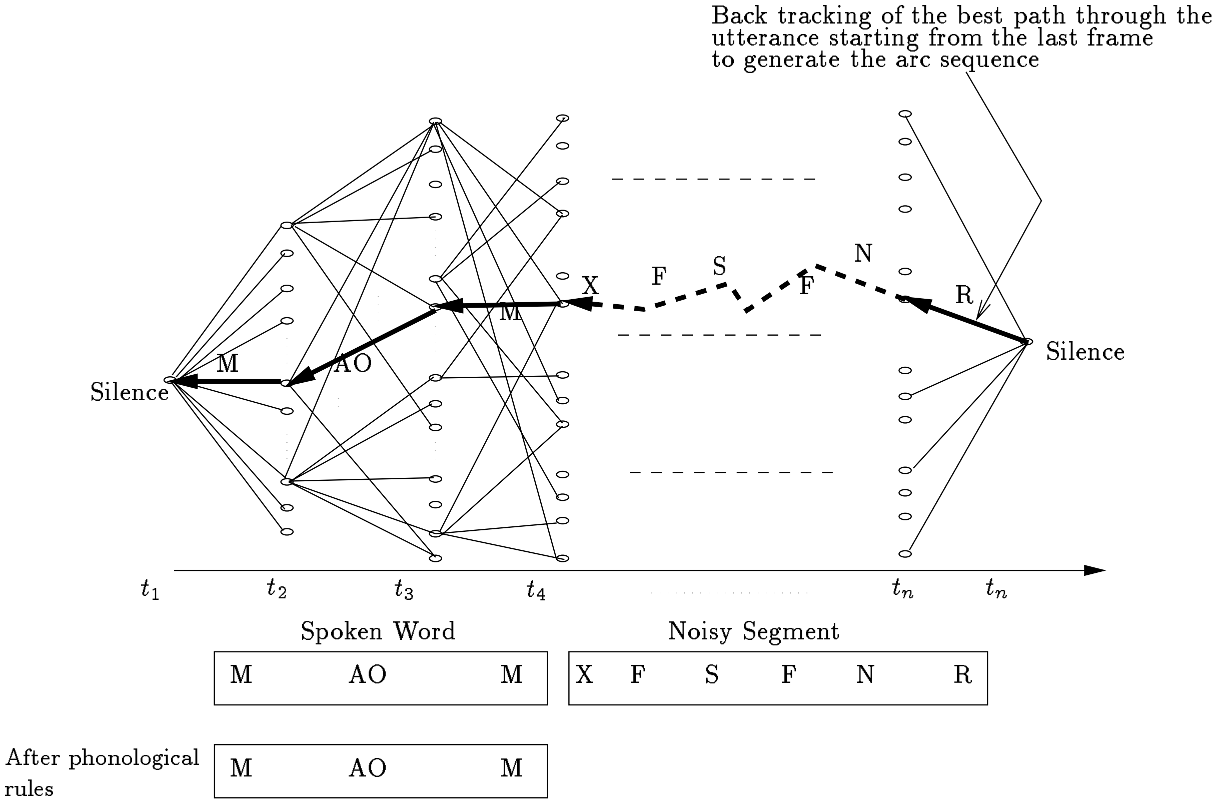

Figure 1: Trellis construction and the generation of a baseform for a sample utterance.

Input Phone Sequence	Replacement Phone Sequence
Silence : consonant sequence (no vowels) : Silence	Silence
Glottal Stops (voiceless stops) : Silence, Fricatives and Nasals : End of Utterance	Glottal Stop
Nasals : Glottal Stops : End of Utterance	Nasals
Vowel : R or B : Vowel	Vowel : V : Vowel

Table 2: Sample Phonological Rules Used to obtain cleaner baseforms

In the next experiment, the acoustic models used for decoding and generation of arc-to-arc transition probabilities used in the ballistic labeler were trained from a large general-purpose training corpus. This included digits, short sentences, spontaneous speech, yellow pages categories, street addresses, and read speech. Less than one percent of the training corpus contained name-specific data, as opposed to the previous experiment where the models were trained purely on name-specific data. DB II was used for experimenting with the baseforms generated after application of phonological rules. The results are tabulated in Table 3. The first column shows the recognition accuracy obtained when no post processing was done. The second column shows the accuracy obtained after application of phonological rules to the baseforms obtained from the ballistic labeler. It can be seen that these generalized models, degrade the quality of the baseforms produced considerably when compared to the domain-specific models. The application of phonological rules to these baseforms help in improving the recognition accuracy.

Experiments are currently in progress for a larger system with 1000 names and 50 speakers averaging 20 utterances of a name. This data was collected in a real-office environment during daily use.

Method Employed	Recognition Accuracy
Ballistic Labeler	89.3%
Phonological Rules	94.7%

Table 3: Recognition performance after post-processing with phonological rules

4. CONCLUSIONS AND FUTURE WORK

The goal of this work was to automatically determine phonetic baseforms for words not used in the recognizer and evaluate the performance of the current speech recognition system for those words. Visual inspection of the baseforms generated, indicated substantial discrepancies between the automatically generated baseforms using acoustics alone and hand-written ones for almost one-tenth of the data. Despite this, the recognition accuracy seemed to be better when using these baseforms as opposed to the ones generated by a linguist. Most of the errors were made in situations where the telephone channel conditions were really noisy, with several people talking in the background, or when the user pronounced the word very differently from call to call. As expected, it was observed that if baseforms were generated with two utterances, both from clean and noisy scenarios, the recognition accuracies improved slightly. Our algorithm produced the following baseform for the utterance 'SHARON KEN', D$ SH EH R AE N K EH N X S X D$ as opposed to the hand written version which had two pronunciations, namely, SH AE R AX N K EH N and SH EH R AX N K EH N. It can be seen that a 'S' phone has been introduced into the baseform sequence. Most of the errors observed arise from these sporadic 'S' , 'V', and 'F' phone insertions, particularly under noisy conditions. Our system uses a 54-phone set where D$ and X are the silence phones. These random phones towards the end which captured the noisy portion of the data, where deleted by the application of rules, and this resulted in the baseform, D$ SH EH R AE N K EH N D$.

The good performance of this algorithm can be attributed to the following fact. The training of the HMM's are done at the leaf level even though the trellis computation and Viterbi search is done at the arc level (a third of a phone) and this fine level of detail that is included in the models, accounts for the high performance accuracy. In order for this algorithm to be of any use in a practical application, the generation of baseforms should be done at speeds much less than real time. To save on computations, a trellis of arc nodes is used. However, a trellis of leaf nodes can also be used for the same purpose without considerable loss in accuracy.

Secondly, in the mapping from arc to phone nodes to obtain the final phonetic sequence, linguistic constraints in the form of phonological rules have been applied. These rules eliminate erroneous phone transitions that could have been generated either by the pruning procedure or by the presence of noise in the speech segment.

Experiments using the N best predecessors to generate alternate pronunciations for an utterance are currently under way. This will be explored along with the use of confidence measures using the acoustic scores of each phone to represent noisy segments and to eliminate weird phone transitions that are currently captured by these rules.

In conclusion, we believe that we have a viable technique for generating good phonetic baseforms that give a high decoding accuracy with our speech recognizer. This is particularly useful for our telephony toolkit, where personalized vocabularies are a must. Work is currently under way to employ this algorithm in other components of the speech recognizer, so that phonetic baseforms existing in the system can be adapted to provide improvements in accuracy for spontaneous speech applications.

5. REFERENCES

1. L. R. Bahl, S. Das, P.V. deSouza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M.A. Picheny, J. Powell, "A utomatic Phonetic Baseform Determination", Proc. Speech and Natural Language Workshop, pp. 179-184, June 1990.
2. J. M. Lucassen and R. L. Mercer. "An Information Theoretic Approach to the Automatic Determination of Phonemic Baseforms", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4 2.5.1-42.5.4, 1984.
3. L.R. Bahl et al., "A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition", IEEE Transactions on Speech and Audio Processing, vol. 1, no. 1,pp 59-67. January 1993.
4. R. C. Rose and E. Lleida "Speech Recognition using Automatically Derived Baseforms", pp 1271-1274, ICASSP 1997.
5. R. C. Rose et al., "A User-Configurable System for Voice Label Recognition" ,Proc. Int. Conf. on Spoken Lang. Processing, October 1996.
6. L.R. Bahl et al. "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task." vol 1, pp 41-44, ICASSP 1995.
7. J. Elvira, J. Torrecilla "Name Dialing using Final User Defined Vocabularies in Mobile (GSM, TACS) and Fixed Telephone Networks", ICASSP 1998.
8. B. Ramabhadran et al. "Acoustics-Only Based Automatic Phonetic Baseform Generation", ICASSP 1998.