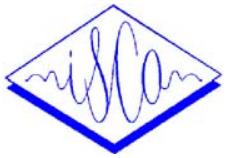

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ACOUSTICS-BASED BASEFORM GENERATION WITH PRONUNCIATION AND/OR PHONOTACTIC MODELS

Bhuvana Ramabhadran, Sabine Deligne, Abraham Ittycheriah

IBM T. J. Watson Research Center P. O. Box 218,
Yorktown Heights, NY 10598
(bhuvana,deligne,abei)@us.ibm.com

10.21437/Eurospeech.1999-130

ABSTRACT

In this paper, we describe a method to derive a phonetic pronunciation of a word using only an acoustic utterance of that word without a priori knowledge of the spelling of the word. In [5] and [6], we used a pronunciation model based on bigram statistics. Bi-gram statistics only constrain the left neighbor phone and results in phone sequences that are only pairwise appropriate. Here, we apply a pronunciation model in combination with a phonotactic model that serves the purpose of a language model to constrain the phone sequences produced. Error rates with and without the phonotactic model are presented.

1. INTRODUCTION

There has been considerable interest in telecommunications and embedded speech recognition applications that provide personalized vocabularies [1] [2] [3]. 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 words to his/her personalized vocabulary, for which a spelling or acoustic representation does not exist in the speech recognition lexicon, and associate these words to a phone number to be dialed. Once the personalized vocabulary is configured, the user can subsequently dial the phone number by speaking the words just added to the vocabulary. In order to do this, proper derivation of the phonetic baseform is essential.
In [5] we showed how speaker dependent baseforms could be derived from one or two speech utterances by using speaker independent acoustic models and *a priori* knowledge expressed by bigram probabilities to constrain the transition between the models. Due to the short span of the bigram constraints, the baseforms produced by this method evidenced phone sequences which were often in contradiction with the phonotactics of the task. Typically, if the enrollment is done during noisy conditions, fricatives and stops are inserted amidst legal phonetic sequences. In [6] we showed how filtering these baseforms based on phonological rules and confidence scores could improve the accuracy of the system.
In this paper we explore the use of various models such as pronounciation and/or phonotactic models to constrain the transitions between the phones during the derivation of the baseforms. The structure of this paper is as follows. In section 2, we first formulate a general framework which ecompass the diversity of these modeling assumptions, and we then propose possible strategies to integrate pronounciation and phonotactic knowledge within this framework. In section 3, we present the data and the experimental protocol used to asses the relative efficiencies of these strategies. Results are discussed in section 4, and section 5 provides extensions and applications for this work.

2. FRAMEWORK FOR THE DERIVATION OF BASEFORMS

2.1. Formulation

The problem of deriving a baseform from acoustic evidence alone can be stated as the problem of retrieving the most likely string P^* of phones, given a string O of T acoustic observations:

$$\begin{aligned} P^* &= \arg\max_{\{P\}} L(P \mid O) \\ &= \arg\max_{\{P\}} L(O \mid P)\, L(P) \end{aligned}$$

where $L(P)$ is the likelihood of a phone string, and where $L(O \mid P)$, the conditional likelihood of acoustic observations, is computed using acoustic models. Assuming the acoustic models are Hidden Markov Models (HMM), there are multiple sequences S of T states in the HMM which can account for the T acoustic observations and result in the same phone string P:

$$P^* = \arg\max_{\{P\}} \sum_{\{S\}} L(O,\ S \mid P)\, L(P)$$

By approximating the likelihood of a phone string with the likelihood of the best underlying sequence of states, and then using the Bayes rule:

$$\begin{aligned} \tilde{P}^* &= \arg\max_{\{P\}} \max_{\{S\}} L(O,\ S \mid P)\, L(P) \\ &= \arg\max_{\{P\}} \max_{\{S\}} L(O \mid S,\ P)\, L(S \mid P)\, L(P) \end{aligned}$$

$$= \arg\max_{\{P\}} \max_{\{S\}} L(O \mid S)\, L(S \mid P)\, L(P) \quad (1)$$

where we have further assumed that $L(O \mid S,\ P)$ equals $L(O \mid S)$, i.e. that the probability of an observation is conditioned by the emitting states only. In Eq. (1), the term $L(O \mid S)$ is determined by the acoustic model, while $L(S \mid P)$ acts as a pronounciation model and $L(P)$ as a phonotactic model. We proceed by explaining in more details the estimation of each of these 3 components.

2.2. Acoustic model

The acoustic component refers to the model used to compute the state conditional likelihood of each observation:

$$L(O \mid S) = \prod_{t=1}^{t=T} p(o_{(t)} \mid s_{(t)})$$

The conditional probability of an observation $p(o \mid s_i)$ is computed with a speaker independent mixture of gaussians. Each mixture models a context dependent subphone unit called a leaf. The phonetic context specified by a leaf can span over up to 5 preceding phones, and the most relevant partition of contexts for each phone is determined by using a decision tree [4]. In our experiments, the states s_i correspond either to the context dependent leaves, or to context independent subphone units called arcs (a phone consists of 3 arcs, corresponding to the three states of a HMM ; for eg., the phone AA is made of arcs, AA-1, AA-2 and AA-3). In a scheme where each state s_i corresponds to a leaf, the probability $p(o \mid s_i)$ is the emission probability of observation o given the leaf pointed by s_i. In a scheme where each state s_i corresponds to an arc, the probability $p(o \mid s_i)$ is taken as the highest emission probability among the set of leaves matching this arc.

2.3. Pronounciation Model

The pronounciation component refers to the model used to compute $L(S|P)$ the likelihood of a state sequence S in Eq. (1). Since in our system, each subphone unit (be it context dependent or not) is modeled with a distinct unique state, any phone string P determines uniquely the identity of the successive states in S. Therefore, what $L(S|P)$ would essentially model is the duration of each of the states forming S. But because in the application at hand the phone strings P are baseforms which are derived precisely because they are missing from the lexicon (e.g. proper nouns), we do not have training data to estimate $L(S \mid P)$. We therefore have to assume that the probability of the current state does not depend on the entire phone string P. In our experiments, we make the assumption that $L(S \mid P)$ equals $L(S)$. In this paper, we assume that each state depends on the previous one only:

$$L(S \mid P) = \prod_{t=1}^{t=T} p(s_{(t)} \mid s_{(t-1)})$$

and we explore 2 schemes to estimate the probabilities $p(s_{(t)}|s_{(t-1)})$. In one scheme, we use bigram transition probabilities $p(s_j|s_i)$ as probabilities of going from the preceding state to the current one. In the other scheme, the pronounciation model is merely a state duration model, and $p(s_{(t)}|s_{(t-1)})$ equals either the probability $p(s_i|s_i)$ of staying in state s_i, or the probability $\sum_{j \neq i} p(s_j|s_i)$ of leaving state s_i. In both schemes, a state can refer to either an arc or a leaf; altogether a pronounciation model can thus be either an arc duration model, an arc bigram model, a leaf duration model or a leaf bigram model.
To compute estimates of the state duration or transition probabilities, a database is aligned to the leaf or arc models and the probability $p(s_j|s_i)$ is computed as the relative count of state s_j in the alignment (no backoff is used).

2.4. Phonotactic model

The phonotactic component refers to the model used to compute $L(P)$ the likelihood of a phone string P in Eq. (1). We assume n-gram dependencies between the phones and we compute the n-gram estimates on a database consisting of phonetic transcriptions (the phonetic transcriptions of the database aligned to collect state transitions statistics, cf. section 2.3) using a backoff strategy. By contrast with the pronounciation model modeling the state transitions, the phonotactic model does not model the duration of the phones, and it is estimated so as to minimize the phone perplexity of the training data. In a previous work [6], we presented experiments where no phonotactic model was used, which is equivalent to assuming that all phone strings are equally likely. In this paper, we compare different strategies with and without phonotactic models.

2.5. Algorithm

Following the assumptions introduced in the previous sections, Eq. (1) is rewritten as:

$$\tilde{P}^* = \arg\max_{\{P\}} \max_{\{S\}} L(O \mid S)\, L(S)\, L(P) \quad (2)$$

Eq. (2) is implemented with a Viterbi algorithm by using a bigram model to compute $L(P)$ (a phone bigram probability is applied only when moving from the last state of a phone to the first state of another phone). All hypotheses in the lattices are constrained to start and end with respectively the first state and the last state of the silence model. Distinct lattices are grown, each of which contains hypotheses of baseforms - partially decoded utterances - consisting of the same number of distinct phones. The scores of the hypotheses are normalized by their number of phones, and the one with the highest score is backtracked as being the retrieved baseform.

3. MODELS, DATA AND PROTOCOL

3.1. Models

The system comprises speaker independent acoustic models for 52 English phones (including a silence model), each

phone being modeled as a concatenation of 3 arcs (subphone units), it has 156 arcs, and there are a total of 2448 leaves (a leaf is a context dependent arc). The mixtures of gaussians modeling each leaf are estimated on a telephony database consisting of digits, names and addresses. A similar database of about 17500 utterances is aligned to the acoustic models to collect the statistics required to estimate the state duration and state transition probabilities of the pronounciation model. The bigram phonotactic model is estimated on the phonetic transcription of the same telephony data (637K phones).

3.2. Enrollment and test data

Our experiments of baseform generation are run for a name-dialing task, using an in-house data collection software for telephony data collection. The database was built using 6 speakers with a variety of accents and each speaker asked to enroll 50 different items, that included names, such as, ROSE, ANTHONY FABRIZIO, DAD, MOM, etc.), and some command words, such as, 'HELP', 'CANCEL', 'CALL RETURN'.
The data were recorded at a 8kHz sampling rate, and 39 dimensional vectors comprising 12 cepstra + the energy, delta and delta-delta coefficients were computed from 10 ms frames.
The starting and ending frames of the speech utterances were automatically detected using 2 sets of gaussians previously estimated on a large database to model respectively speech and silence. The reason for using speech detection is that the baseform derivation algorithm is likely to produce spurious sequences of phones in place of the beginning and ending silent portions, which would increase the confusability between the baseforms across the vocabulary.

3.3. Protocol

For each speaker, baseforms were generated from 2 speech utterances and the remaining calls were used as the test bed for recognition with the newly generated baseforms. Before being added to each speaker's lexicon, the derived baseforms are automatically filtered using phonological rules, so that sequences of phones which are likely to be spurious sequences, such as consonnants intermixed with silences, be removed (this acts as a complement to the speech detection step).
As a baseline experiment, we performed recognition on the test set using hand-written baseforms as defined by a linguist. Then in order to assess the relative efficiencies of the pronounciation and phonotactic models, 3 sets of experiments are presented. In a first set of experiments, it is assumed that all phone transitions are equally likely (no phonotactic model) and various pronounciation models are compared (arc duration, arc bigram transitions, leaf duration and leaf bigram transition). In a second experiment, it is assumed that all state transitions are equally likely (no pronounciation model) and a bigram phonotactic model is used. In a third set of experiments, the experiments of the first set where the states refers to arcs are reproduced with the integration of a phonotactic model. Also, all the experiments were done again by suppressing the filtering step. All results are given as word accuracies averaged over all the speakers, and they are discussed in section 4.

4. RESULTS AND DISCUSSION

Hand-written baseforms:

As a baseline experiment, the set of utterances used as a test set was decoded using the hand-written baseforms. The average word accuracy obtained across all the speakers is 97.5%.

Baseforms derived using various pronounciation models, without phonotactic model:

The average word accuracy obtained across all the speakers for these experiments are in Table 1. As could be expected, the bigram models outperform the duration models, and the models at the leaf level outperform the models at the arc level. Indeed, while the identity of an arc informs only on the identity of the current phone, the identity of a leaf implicitly conveys contextual information relating to the 5 preceding phones, so that a transition model at the leaf level is far more specific than a transition model at the arc level. The performance of the baseforms derived with the leaf model equals the performance of the hand-written baseforms. However the difference in performance between the leaf and arc bigram models is not significant, especially when considering the increase of algorithmic complexity resulting from the use of the leaf model (2448 leaves candidates at each time frame versus 156 arc candidates).

	arc	leaf
duration model	93.3	95.9
bigram model	97.2	97.4

Table 1: Word accuracy with baseforms derived using various pronounciation models, without phonotactic model

Baseforms derived without pronounciation models, using a bigram phonotactic model:

The average word accuracy obtained across all the speakers is 89.6%, which is to be compared with the 97.2% accuracy yield by the arc bigram model The gap between the two scores reflects the importance of modeling the duration of each state within the phone units and not only the transitions from one to another phone.

Baseforms derived using various pronounciation models at the arc level, and a bigram phonotactic model:

The average word accuracy obtained across all the speakers for these experiments are in Table 2. In our experiments, integrating a bigram phonotactic model with a pronounciation model degrades the word accuracy. Proper nouns are precisely those nouns the phonotactics of which tend to be unpredictable (especially as the vocabulary of our experiments include names from different countries). In the case where only a pronounciation model is used, the impact of the phone transition estimates is lessened by the phone duration estimates. The integration of a phonotactic model enhances the impact of the phone transition estimates with respect to the phone duration estimates. Our results seem to indicate that our phone transition estimates are not so robust that there is any benefit in enhancing their impact. Another issue is how to integrate the phonotactic model together with the other components: this issue is similar to the problem of integrating a language model component together with the acoustic component of a recognition system, where usually a language model penalty is used to not disqualify sentences with many words. In our experiments, we have normalized the scores of all the hypothesized baseforms by their number of phones[1], which may not be an optimal strategy.

arc duration model	88.9
arc bigram model	95.9

Table 2: Word accuracy with baseforms derived using pronounciation models at the arc level, and a bigram phonotactic model

Finally, all above experiments were repeated by suppressing the filtering of the baseforms. It slightly degraded the scores of the first set of experiments (where no phonotactic model is used) but did not affect the scores of the second set of experiments (with a phonotactic model), meaning that the effect of the phonotactic model encompass the effect of the filtering. The filtering of the baseforms is a rule based approach to constrain the phone sequences, as opposed to the statistical approach of the phonotactic model.

5. CONCLUSIONS AND PERSPECTIVES

We have presented a general framework for the automatic derivation of baseforms. Assuming a fixed set of acoustic models of subphone units, we have explored various strategies to best constrain the transitions between the subphone units during the derivation. In our experiments, integrating *a priori* knowledge of the duration of the subphone units seems essential. Enhancing the impact of phonotactic constraints reduces the insertions of spurious phones and thus eliminates the need of post-processing the baseforms with phonological rules. However for our task which essentially deals with proper names, it did not bring any improvement. As the phonotactics of proper names are known to be difficult to predict, larger databases should be used in order to compute more reliable estimates. Besides the optimal way to integrate the phonotactic constraints together with the other components remains an issue. Also, at present, alternate pronunciations (baseforms) are generated by using 2 utterances during enrollment. Instead, N best lists could be used to generate lattices of multiple pronunciations for each utterance.

In conclusion, we have a technique for generating phonetic baseforms that give a decoding accuracy similar to the one obtained with the hand-written baseforms with our speech recognizer. This is particularly useful for our telephony toolkit, and for embedded applications in general, where personalized vocabularies are a must.

[1] Note that the scores obtained without applying this normalization were not dramatically different.

6. 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] R. C. Rose et al., "A User-Configurable System for Voice Label Recognition" ,Proc. Int. Conf. on Spoken Lang. Processing, October 1996.

[3] R. C. Rose and E. Lleida "Speech Recognition using Automatically Derived Baseforms", pp 1271-1274, ICASSP 1997.

[4] 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.

[5] B. Ramabhadran, A. Ittycheriah. "Phonological Rules for Enhancing Acoustic Enrollment of Unknown Words". ICSLP 98.

[6] B. Ramabhadran, L.R. Bahl, P.V. DeSouza, M. Padmanabhan. "Acoustics-Only Based Automatic Phonetic Baseform Generation", ICASSP 98. 2.