AUTOMATIC GENERATION AND SELECTION OF MULTIPLE PRONUNCIATIONS FOR DYNAMIC VOCABULARIES

Sabine Deligne, Benoit Maison, and Ramesh Gopinath

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

ABSTRACT

In this paper, we present a new scheme for the acoustic modeling of speech recognition applications requiring dynamic vocabularies. It applies especially to the acoustic modeling of out-of-vocabulary words which need to be added to a recognition lexicon based on the observation of a few (say one or two) speech utterances of these words. Standard approaches to this problem derive a single pronunciation from each speech utterance by combining acoustic and phone transition scores. In our scheme, multiple pronunciations are generated from each speech utterance of a word to enroll by varying the relative weights assigned to the acoustic and phone transition models. In our experiments, the use of these multiple baseforms dramatically outperforms the standard approach with a relative decrease of the word error rate ranging from 20% to 40% on all our test sets.

1. MOTIVATION

Speech recognition systems usually rely on a fixed lexicon where the pronunciations of the vocabulary words are given by hand-crafted phonetic baseforms, i.e. sequences of phones written by a phonetician. However, many applications require new words to be dynamically added to the recognition vocabulary, or new pronunciations of invocabulary words to be added to the lexicon. Hence the need for techniques which can automatically derive phonetic baseforms. This occurs for example in dictation systems that allow personalized vocabularies, in name dialer applications where the user enrolls the names he wants to dial, and in any application where actual pronunciations differ from canonic pronunciations (like for non-native speakers), so that the robustness of linguist-written pronunciations needs to be improved. In situations where the speech recognition engine is embedded in a small device, there may not be any interface media, such as a keyboard, to allow the user to enter the spelling of the words he wants to add to his/her personalized vocabulary [1]. And even if such an interface were to be available, the spellings may not be of very much help as these applications typically involve words the pronunciation of which is highly unpredictable, like proper names for example. In this context, it is difficult to use *a priori* knowledge, such as letter-to-sound rules in a reliable way. Consequently, the user is asked to utter once or twice the words to add to his/her personalized vocabulary, and phonetic baseforms for these words are derived from the acoustic evidence provided by the user's utterances. These approaches ([2], [3], [4], [5]) usually rely on the combined use of: (i) an existing set of speaker-independent acoustic models of subphone units, and (ii) a model of transition between these subphone units. The way to optimally combine these models is an open issue as it is not known in advance which of the models can most reliably describe the acoustic evidence observed for each new word to enroll. For example, when the enrolled words are proper names, the reliability of the model of transition between the subphones is questionable since proper names do not follow strict phonotactic rules. Current techniques of automatic baseform generation do not take into consideration the relative degree of confidence that should be put in either component. The scheme presented in this paper deviates from standard approaches in that: (i) the acoustic model and the transition model which are combined to generate the baseforms are assigned a weight, (ii) multiple baseforms are derived from a single speech utterance by varying the relative weights of the models. The basic idea behind this approach is twofold. First, since we have to guess the pronunciation of the enrolled words from just one or two speech examples, we may as well use multiple guesses to maximize the chance of guessing right. Second, since we do not know *a priori* how reliable each of the two models is relative to the other model, we avoid arbitrarily favoring either one of the models by varying their relative weights when generating the guesses. The distinct baseforms obtained from the speech utterance of a word are added to the recognition lexicon as pronunciation variants of that word. It has been extensively investigated recently how, in standard recognition frameworks, adding pronunciation variants to the canonic pronunciations of a static lexicon can significantly improve the

recognition accuracy [7] [8]. We show that this applies also in the context of dynamic vocabularies, where no canonic pronunciations at all are available. On the other hand, as multiple baseforms are added to the recognition lexicon, we can expect the acoustic confusability between the entries of the lexicon to increase with the risk of hurting the recognition accuracy. In this paper, we report on an extensive set of speech recognition experiments showing the influence of the number of automatically generated baseforms on the decoding accuracy. The structure of this paper is as follows. In section 2 and 3, we describe our scheme to generate multiple baseforms and build variable-size lexicons. In section 4, we present speech recognition experiments comparing lexicons of automatically generated baseforms on test data consisting of either isolated or in-context words, and in both quiet and noisy environments. Section 5 concludes on this work.

2. GENERATION OF MULTIPLE BASEFORMS

In this section, we present a scheme to derive multiple baseforms from acoustic evidence, where it is attempted to make the best possible use of our *a priori* knowledge, where our *a priori* knowledge comprises a set of speaker independent acoustic models of subphone units and a statistical model of transitions between subphone units[1]. In the standard way, the problem of deriving a baseform from acoustic evidence is usually stated as the problem of retrieving the most likely string U^* of T subphone units, given the string O of T acoustic observations:

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

The string U^* is retrieved with a Viterbi algorithm. The conditional probability of the acoustic observations given the string of subphone units is computed as:

$$P(O \mid U) = \prod_{t=1}^{t=T} p(o_{(t)} \mid u_{(t)})$$

The conditional probability of each acoustic observation- $p(o \mid u_i)$ is computed with the acoustic model - in our experiments speaker independent mixtures of gaussians. The probability of observing the string of subphone units U is computed with the transition model assumed between the subphones - in our experiments a bigram model:

$$P(U) = p(u_{(1)}) \prod_{t=2}^{t=T} p(u_{(t)} \mid u_{(t-1)})$$

[1] Each subphone unit corresponds to roughly one third of a phone.

The bigram model of subphone units is estimated off-line by aligning a large dataset of speech with a known transcription on the acoustic models of the subphone units. The probabilities $\{p(u_j \mid u_i)\}$ are computed from the observed relative counts of the subphone models in the alignment (in our experiments, no backoff is used to smooth the bigram probabilities). Note that both the duration of the units and the transition between the units are modeled.
The modification that we introduce to this baseline approach is to compute the log-likelihood of a baseform as a weighted sum of the log-scores of the acoustic model and of the transition model, with weights respectively of $(1-\lambda)$ and λ:

$$U^*_\lambda = \arg\max_{\{U\}} (1-\lambda) \log P(O \mid U) \; + \; \lambda \log P(U)$$

Each value of λ defines a distinct log-likelihood function wich reaches its maximum value for possibly distinct strings of subphone units. The parameter λ can be seen as reflecting the confidence put into each model. In a context where it is not known which of the two models can most relevantly account for the observations, the generation of multiple strings U^*_λ for various values of λ allows to compensate for a possible mismatch.

3. BUILDING LEXICONS WITH MULTIPLE BASEFORMS

In our experiments, we define a set of λ values by scanning an interval $[\lambda 1; \lambda 2]$ ($0 \leq \lambda 1 \leq \lambda 2 \leq 1$), with a step of 0.1. Each string U^*_λ is converted into a phonetic baseform by replacing the subphone units with their phone counterpart and by merging together repeated phones. All the distinct phonetic baseforms obtained from the speech utterance of a word by scanning a set of values of λ are added as pronunciation variants in the recognition lexicon. Each interval $[\lambda 1; \lambda 2]$ thus results in a specific recognition lexicon, hence raising the question of how to select *a priori* the best performing lexicon. We can expect that accumulating multiple baseforms for each enrollment speech utterance will improve the recognition accuracy by allowing a broader modelling of the pronunciation of the new words. However it is well known that increasing the number of pronunciation variants increases the acoustic confusability in the recognition lexicon, which eventually hurts the accuracy. In our experiments, we noticed for example that the baseforms obtained with λ equal to or more than 0.8 tended to look more and more alike, which we attributed to the prevailing influence of the subphone transition model. As a result, cumulating baseforms with λ values higher than 0.8 was resulting in higher word error rates. In the following section, we report on experiments where lexicons are build for each interval $[\lambda 1; \lambda 2]$ with $\lambda 1$ in $\{0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7\}$ and $\lambda 2$ in $\{\lambda 1; ...; 0.7\}$. The selection of the most promising

lexicon thus resumes to selecting the appropriate interval $[\lambda 1; \lambda 2]$.

4. EXPERIMENTS

4.1. Baseform generation

We report on experiments with 2 different sets of enrolled words: (i) the enrollment set $E1$ consists of 50 distinct words, each word being repeated twice by 10 speakers, (ii) the enrollment set $E2$ consists of 35 distinct words, each word being repeated once by 20 speakers. All the data are recorded using a push-to-talk button in a quiet environment at 22kHz and downsampled to 11kHz. The front end computes 12 cepstra + the energy + delta and delta-delta coefficients from 15ms frames. Baseforms are automatically generated using a reduced-size acoustic model especially designed to be used in portable devices, or in automotive applications. It consists of a set of speaker-independent acoustic models (156 subphones covering the phonetics of English) with about 9,000 context-dependent gaussians (triphone contexts tied by using a decision tree [9]), trained on a few hundred hours of general English speech (about half of these training data has either digitally added car noise, or was recorded in a moving car at 30 and 60 mph). The bigram model of subphones was estimated off-line on an aligned corpus of about 17,000 sentences (essentially names, addresses, digits). Speaker-dependent lexicons $E1_{\lambda 1,\lambda 2}(S)$ and $E2_{\lambda 1,\lambda 2}(S)$ are formed for each speaker S in respectively $E1$ and $E2$, following the procedure described in section 3.

4.2. Baseform evaluation

The recognition lexicons $E1_{\lambda 1,\lambda 2}(S)$ derived for each speaker in the enrollment set $E1$ are evaluated on 2 test sets: (i) the test set $T1.1$ where each of the 50 words in $E1$ are repeated in isolation 10 times by each of the same 10 speakers, (ii) the test set $T1.2$ where each of the 50 words in $E1$ are repeated in 10 different short sentences (typically command sentences like "ADD $< name >$ TO THE LIST", where $< name >$ is an enrolled word) by each of the same 10 speakers. The recognition lexicons $E2_{\lambda 1,\lambda 2}(S)$ derived for each speaker in the enrollment set $E2$ are evaluated on 3 test sets: (i) the test set $T2.1$ is recorded in a quiet environment, (ii) the test set $T2.2$ is recorded in a car moving at $30mph$, (iii) the test set $T2.3$ is recorded in a car moving at $60mph$. All 3 sets $T2.1$, $T2.2$ and $T2.3$ consist of the 35 words in $E2$ uttered once and preceded by either the word "CALL", "DIAL" or "EMAIL", by each of the speakers in $E2$. The baseforms of the command words "CALL", "DIAL"... in the test sets are linguist-written baseforms.

4.3. Recognition scores

Figure 1 plots the Word Error Rate as a function of the interval $[\lambda 1; \lambda 2]$: the points along the x axis correspond to the intervals $[0.1; 0.1]$, $[0.1; 0.2]$, ... ,$[0.1; 0.7]$, ..., ending with the intervals $[0.6; 0.6]$, $[0.6; 0.7]$ and $[0.7; 0.7]$. The WER corresponding to a standard generation system ($\lambda 1 = \lambda 2 = 0.5$) are circled. The solid line represents the WER averaged over all the speakers in both test sets $T1.1$ and $T1.2$, i.e. decoding with the lexicons $E1_{\lambda 1,\lambda 2}(S)$. The dot line represents the WER averaged over all the speakers in the test sets $T2.1$, $T2.2$ and $T2.3$, i.e. decoding with the lexicons $E2_{\lambda 1,\lambda 2}(S)$. As can be seen, both curves show the same local patterns: the WER decreases along each portion of the x axis going from an interval $[\lambda 1; \lambda 1]$ to an interval $[\lambda 1; 0.7]$, which indicates how accumulating baseforms systematically improves the overall accuracy. Also, the general pattern of both WER curves is to increase towards the ending intervals, intervals starting with a $\lambda 1$ more than 0.5. The curves on Figure 1 tend to indicate that a close-to-optimal accuracy can be obtained by building a lexicon using an interval $[\lambda 1; 0.7]$, where $\lambda 1 \leq 0.3$.

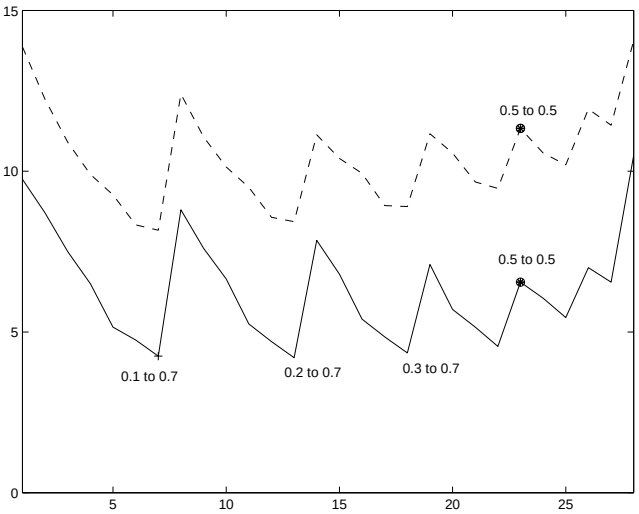

Fig. 1. WER as function of the chosen interval $[\lambda 1; \lambda 2]$ defining the set of values of λ, averaged on the set $T1.1$ and $T1.2$ (solid line), and , averaged on the set $T2.1$, $T2.2$ and $T2.3$ (dot line).

In Table 1, we show WERs on $T1.1$ and on $T1.2$, (averaged over all speakers): WER(worst) and WER(best) are respectively the highest and the lowest WER over all intervals $[\lambda 1; \lambda 2]$; WER(standard) is the standard WER obtained with $\lambda 1 = \lambda 2 = 0.5$; and WER(0.1,0.7) is the WER obtained with the interval $[0.1; 0.7]$. Table 2 shows the same statistics on the sets $T2.1$, $T2.2$ and $T2.3$. In both tables, decoding with the lexicon obtained by scanning the interval $[0.1; 0.7]$ usually equals the best performance that can

	T1.1	T1.2
WER(worst)	8.8	10.1
WER(standard)	7.2	5.9
WER(best)	4.4	3.9
WER(0.1,0.7)	4.5	3.9

Table 1. WER on $T1.1$ and $T1.2$ with four lexicons selected among all $E1_{\lambda 1,\lambda 2}$

	T2.1	T2.2	T2.3
WER(worst)	13.1	12.9	16.3
WER(standard)	10.5	10.6	12.9
WER(best)	7.3	6.7	10.3
WER(0.1,0.7)	7.3	6.7	10.5

Table 2. WER on $T2.1$, $T2.2$ and $T2.3$ with four lexicons selected among all $E2_{\lambda 1,\lambda 2}$

be obtained over all the lexicons. Besides, it yields a relative WER improvement of more than 30% over the standard approach on all test sets, except on the set $T2.3$ (test data recorded in a car at 60mph) where the relative decrease of the WER is 18%. Note that the WER could be further decreased by pre-processing the speech data with speech detection, and by post-processing the baseforms by automatically filtering out the suspicious sequences of phones. Indeed, these two techniques were shown to improve the recognition accuracy [6].

5. CONCLUSIONS AND PERSPECTIVES

We have introduced a simple, but extremely powerful (from an accuracy viewpoint) scheme to derive the pronunciation of new words in dynamic vocabularies. Pronunciation variants of the words are automatically generated from a few enrollment speech utterances. The variants are obtained by varying the relative weights of the acoustic and transition models combined to retrieve the pronunciation. In our experiments, this approach yielded a relative decrease of the word error rate ranging from 20% to 40% on all our test sets.

One issue raised by this technique is the choice of the appropriate range of weight values over which to accumulate the pronunciations. In our experiments, the optimal range, for a given acoustic and transition model, turns out to be very stable across all test sets. It would however be desirable to automatically detect the point where the benefit of adding more pronunciation variants to the lexicon no longer compensates the increase of the acoustic confusability. Measures of acoustic confusability have been increasingly investigated lately, with a special focus on how to estimate a Word Error Rate on test data without having to decode them. We are currently investigating the use of the the Synthetic Acoustic Word Error Rate (SAWER) introduced in [10], which fuses information from both a language and an acoustic model.

6. REFERENCES

[1] L. R. Bahl, S. Das, P.V. deSouza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M.A. Picheny and J. Powell, "Automatic Phonetic Baseform Determination", Proc. Speech and Natural Language Workshop 1990, pp. 179-184.

[2] R. C. Rose et al., "A User-Configurable System for Voice Label Recognition", ICSLP 1996.

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

[4] B. Ramabhadran and A. Ittycheriah, "Phonological Rules for Enhancing Acoustic Enrollment of Unknown Words". ICSLP 1998.

[5] B. Ramabhadran, L.R. Bahl, P.V. DeSouza and M. Padmanabhan, "Acoustics-Only Based Automatic Phonetic Baseform Generation", ICASSP 1998.

[6] B. Ramabhadran, S. Deligne and A. Ittycheriah, "Acoustics-Based Baseform Generation with Pronunciation and/or Phonotactic Models", EUROSPEECH 1999.

[7] H. Strik and C. Cucchiarini (1999), "Modeling pronunciation variation for ASR: A survey of the literature". Speech Communication 29, pp 225-246.

[8] J.E. Fosler-Lussier (1999), "Dynamic pronunciation models for automatic speech recognition", Ph.D. thesis, University of California, Berkeley.

[9] L.R. Bahl et al., "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task", ICASSP 1995, vol.1, pp 41-44.

[10] H. Printz and P. Olsen, "Theory and practice of acoustic confusability", ASR 2000, pp 77-84.