VERY LARGE VOCABULARY PROPER NAME RECOGNITION FOR DIRECTORY ASSISTANCE

Frédéric Béchet, Renato de Mori, Gérard Subsol

LIA - University of Avignon, France
{frederic.bechet,renato.demori,gerard.subsol}@lia.univ-avignon.fr

ABSTRACT

This paper deals with the difficult task of recognition of a large vocabulary of proper names in a directory assistance application. After a presentation of the related work, it introduces a methodology for rescoring the N-best hypotheses generated by a first step recognition. First experiments give encouraging results and several topics for future research are presented.

1. INTRODUCTION

Recognition of a large vocabulary of proper names is a difficult task of a very high perplexity. Taking into account all possible distortions of proper names may lead to an even larger number of word models which are difficult to derive. Furthermore, a given speaker may introduce only certain types of distortions with respect to a canonical representation of words. Thus, search based on a network with all possible distortions of canonical forms may lead to an increase in word error rate because the knowledge used includes a large number of distortion models which are inconsistent with the distortion types introduced by a given speaker.

Another important problem arises from the difficulty to distinguish the fact that a speaker does not use the supposed canonical form of a word from the fact that certain phonemes really uttered by a speaker are not scored by a recognizer with an adequate score. Nevertheless, different pronunciations of a word are very often perceived to be similar, if not equal to the canonical form for that word.

This suggests to guide recognition by canonical pronunciations with the possibility of dynamically allowing plausible deviations from it, rather than using an enriched lexicon with all possible alternate word pronunciations. This can be conveniently performed in complex applications, like Directory Assistance (DA), in various steps in which lattices of hypotheses are progressively reduced until a single hypothesis is selected as recognition result.

This paper introduces a methodology, based on the above considerations, for rescoring the N-best hypotheses generated by a first step recognition, performed after a short dialogue, by a system developed at France Telecom R&D for the whole French directory.

Rescoring is performed by executing a full search on a matrix frame-by-frame acoustic likelihoods of all the phonemes. This is done using a variety of phoneme sequence statistical models and dynamic generation of constrained distortions of the canonical form.

After a discussion on the state of the art in section 2, the theory and application of rescoring with dynamically generated plausible distortions will be presented in section 3 with preliminary results. In section 4, future work to improve the accuracy and the robustness of the rescoring is presented.

2. RELATED WORK

Dealing with the large amount of proper names in a directory assistance application, several specificities can be found that make the problem of pronunciation distortions even more difficult. Major causes for these difficulties and proposed solutions are reviewed in the following.

2.1. Proper names in directory assistance application

- *The inner structure of proper names:* the problem of the automatic speech recognition of proper names is a special one because proper names may have different structure from common words and have special rules for the decomposition into sub-word units [16]. To handle this problem, in [16], a grammar based representation of proper names based on morphs is thus proposed. Morphs are recognized and names are derived from the recognized morphs. A grammar of possible morph sequences representing words is provided and a phoneme representation of each morph is hand-coded.

- *The size of the lexicon:* in [11], it is proposed to build whole word models from phoneme HMMs for representing the 280 surnames that corresponds nearly to all the Korean names. But, in most countries, the set of proper names may include several hundred thousands names (for example, the IBM system described in [9] uses 280,000 names with finite state acceptors representing the combinations of units from a basic vocabulary of 76Kwords). In [9], a claim is made that new recognition methods have to be developed, especially in the field of speaker clustering, massive adaptation of previous calls, unsupervised sentence adaptation, or derivation of personalized vocabularies.

- *Many non-native speakers or words:* there is evidence that current speaker independent recognition systems tend to perform worse when recognizing non-native speech and that this is due to poor acoustic modelling [2]. In [18] the conjecture is made that foreign speakers tend to use phonemes from the native language (source language) to replace phonemes in the target language they do not know how to utter. A mapping between phonemes in the source and the target language can be obtained by forcing alignment of the target model and comparing this with the an initial setting of the mapping which can be further refined by using knowledge from phonetic literature and manual transcription of sentences.

Moreover, it is noted in [9] that non-native speakers are a major source of errors because they do not use the canonical pronunciation and that including new pronunciations in the finite-state name grammars is beneficial.

It is the same for the pronunciation of foreign names by native speakers that may also be highly variable and unpredictable.

- *Unknown orthographic transcription:* if orthographic transcriptions of a proper name is not available (e.g., a name is proposed only by voice to a mobile telephone for recording a phone number), then a baseline word representation has to be determined only from phonetic transcriptions. A possibility could be that of matching the features of an uttered sentence to which a proper name or telephone number is associated with the description of the incoming sentence. This often requires a higher memory capability and has little generalization power. For such a reason in [14] a method is proposed which generates models based on acoustic units.

2.2. High and unpredictable variability of pronunciations

Proper names have a high and unpredictable variability of alternate. A recent survey of the literature on alternate pronunciations can be found in [17]. To deal with this problem, two types of methods are proposed:

1. In [6], it is argued that all possible alternatives to the canonical pronunciations should be available to a recognizer. It is proposed to find them using a Boltzmann machine which has binary inputs corresponding to letters and their context and binary outputs which can be mapped into phoneme symbols corresponding to the most likely pronunciations of the input string.

Some other approaches to automatic learning of alternate pronunciations are described in [4, 15]. The process has essentially four steps : generate alternative, align with reference, derive rules, prune rules. In [13], genetic algorithms are used for determining the letter and the phoneme contexts for text-to-phoneme rule translation. A rule-based approach is proposed in [12] that is based on rule templates.

2. Dynamic lexicons are lexicons based on decision trees with different transcription probabilities for different word context. The lexical representation used for decoding is dynamically determined based on the context. Details can be found in [8] where it is proposed to model the distribution of phone pronunciations jointly at the syllable and word levels. Since phones at syllable boundaries still vary with context, pronunciations in these models include dependencies on the neighboring baseform phone. Other form of context, such as word identity, speaking rate [7], word predictability, are included in the model. A training corpus is needed consisting of a phone recognition transcript aligned to canonical dictionary pronunciation models

In [5], a solution to the open problem of the combination of sub-phone units is proposed. Multiple pronunciations for a name are derived by making vary the weight of a linear combination of logarithms of the probabilities of the acoustic models and the sub-unit models

Moreover, some of the phoneme modifications due to context, such as vowel reduction and phoneme substitution are well captured by triphone or context dependent models provided that enough training data have been used for these models. Syllable deletion, on the contrary requires explicit alternate pronunciations [10].

In practice it has been observed that most (as much as 85% in our experiment) of the phonemes of the canonical form are present in a pronunciation, but they often do not contribute to the sequence of maximum score. Frequent errors for very large vocabularies are observed even if models for alternative pronunciations for each word are used. This is due to several facts, namely:

- Alternate pronunciation models may be incorrect,
- Some alternate pronunciations are not taken into account,
- Recognition errors are not taken into account

3. ALTERNATIVE PRONUNCIATIONS AND RECOGNIZER ERRORS

3.1. Experimental setup

The test corpus consists of 1000 utterances of {first name-last name} from different speakers collected by France Telecom R&D in the frame of its internal directory.

The lexicon consists of the canonical phonetic transcription of 128K different {first name-last name} items.

The first recognition step (baseline system) was performed by the France Telecom recognizer. As a result, it gives for each utterance:

- a N-best list of {first name-last name} items,
- a phoneme lattice that contains the likelihood of the phoneme f_i at frame t, corresponding to the signal part A_t: $P(A_t|f_i)$. The frame step is 16ms.

3.2. Rescoring process

The approach proposed in this paper focuses on the dynamic generation of plausible distortions of canonical forms in a rescoring phase in which the probability of the distortion depends on the nature and the evidence of the competing hypotheses.

In order to analyze the effect of distortion, a simple unigram phoneme model is used for this purpose. Insertion penalties are scored by a distortion probability in a Bayesian decision framework. As this probability is conditioned by a very large population of different events, event descriptions are obtained and broad classes of descriptions are introduced in analogy to the clustering of word histories in language modeling.

Descriptions take into account the shared sequences of phonemes between competing hypotheses because their modifications may have a similar effect on competing hypotheses without affecting the ordering after rescoring.

Search in the lattice of phoneme hypotheses is carried to find the word W (with the canonical phonetic transcription W_c) of the N-best list that maximizes the following posterior probability:

$$P(W \mid A) = \sum_{\tau} P(W_c \tau \mid A) = \frac{1}{P(A)} \left\{ \sum_{\tau} P(A \mid \tau) P(\tau) P(W_c | \tau A) \right\} \quad (1)$$

Where τ is a path in the phoneme lattice, A is a sequence of vectors of acoustic parameters. The recognized word can be obtained by the following decision rule:

$$W^* = \underset{W_c}{\operatorname{argmax}} \left\{ P(A \mid \tau) P(\tau) P(W_c | \tau A) \right\} \quad (2)$$

The rescoring is based on a A* decoding strategy. A unigram phonotactic model was used, obtained with the canonical forms of the entire directory. In addition to the effect of the phonotactic model, this type of rescoring benefits from the possibility of performing an admissible exhaustive search of the best segmentation.

3.2. First experiment: simple rescoring

A first experiment was conducted assuming $P(W_c|\tau A)=1$ and looking for the segmentation of the canonical form of each candidate of the N-best list using the score: $P(A|\tau)P(\tau)$.

Table I summarizes the results obtained by rescoring the 5-best list. The correctness percentage of rank *i* corresponds to the occurrence of the correct name in the first *i* elements of the 5-best. Even with no alternate pronunciation, significant improvement were obtained over the baseline system.

Rank in 5-best	% correct after first pass	% correct after rescoring step
1	500 (50.10%)	585 (58.62%)
2	+110 (61.12%)	+71 (65.73%)
3	+39 (65.03%)	+20 (67.74%)
4	+29 (67.94%)	+12 (68.94%)
5	+14 (69.34%)	+4 (69.34%)

Table 1: Improvements after the rescoring step.

3.3. Second experiment: rescoring with one insertion

The consideration of possible distortions has to proceed by steps in which knowledge is applied to introduce families of plausible phonetic and phonological phenomena.

Phoneme substitution is often taken into account in the mixture of Gaussians used in context-dependent phone models. On the contrary, phoneme insertion, that has a high evidence in the acoustic data, is probably due to a distortion introduced by the speaker. In order to perform a systematic investigation, only one insertion was allowed for a sequence {first-name-last name}.

It is well known, in French, that silence [sil] segments can be inserted between two phonemes (this is frequent for certain pairs of consonants, like [m][n]) and that the vowel [e] can be inserted after a consonant at the end of a utterance.

If a phoneme [y] is inserted between string B and string E, the following probability, according to the equations (1) and (2) should take into account the plausibility of this insertion:

$$P(W_c|\tau A)=P(BE|ByE,A) \quad (3)$$

The probability in (3) should express the evidence that phoneme [y] is inserted because it makes more natural the pronunciation of E after B or is inserted by an error of the recognizer because it represents an intermediate configuration of the features at the end of B and those at the beginning of E.

As it would be impossible in practice to estimate directly from data this probability for every strings B and E and for every A, in analogy with what is done for language modeling, strings ByE are clustered into sequences of a limited number of symbols and A is represented by qualitative descriptors of the acoustic pattern around y.

For the insertion of a silence, the plausibility of inserting between the first phoneme of E and last phoneme of B was represented by two symbols (H:high and L: low). Furthermore, a qualitative evidence of y, in the context of the preceding and the following phoneme was evaluated with the symbol derived from histograms of posterior probabilities. The silence duration was evaluated as short: S and long: L. As not enough data were available for a reliable estimation of the probability in the (3), such a probability was set to 1 when the description corresponded to high plausibility and zero otherwise. The results are reported in Table II.

Rank in 5-best	% correct after first pass	% correct after rescoring step
1	500 (50.10%)	600 (60.12%)
2	+110 (61.12%)	+59 (66.03%)
3	+39 (65.03%)	+18 (67.84%)
4	+29 (67.94%)	+11 (68.94%)
5	+14 (69.34%)	+4 (69.34%)

Table 2: Improvements obtained by rescoring considering [sil] and [e] (sil) insertions.

It appears that, even with simple insertion type, the use of symbols, derived from histograms, to assess plausibility leads to a noticeable improvement.

4. FUTURE WORK

Research continues by exploring ways of improving the accuracy and the robustness of the rescoring process by representing distortion plausibility using the evidence of phonetic features and by integrating a multi-decoder scheme in the first recognition step.

4.1. Representing plausibility by using features evidence

Let be a set of K phonetic features (e.g. the binary feature system described [3]. At each frame t, for the k^{th} feature, it is possible to compute an index of presence φ_{kt} by:

$$x_{kt} \Leftarrow \varphi_{kt} = \frac{\sum_{f \in S_k} P(A_t \mid f)P(f)}{\sum_{\text{all phonemes } f} P(A_t \mid f)P(f)}$$

where S_k is the set of phonemes in frame t having the k^{th} feature.

We can then define a symbol x_{kt} that represents an interval obtained by histogram analysis of the distribution of φ_{kt}. A path τ in the phoneme lattice is then defined by a sequence of frames each of which can be described by a vector X_t of indices of evidence for different phonetic features: $X_t=[x_{1t},\ldots,x_{kt},\ldots,x_{Kt}]$

The plausibility of inserting the phoneme [y] between the phonemes [a] and [b] can be estimated by:

$$P(ab \mid ayb, A) = \prod_k e^{-\gamma(x_{ak}, x_{yk}, x_{bk})}$$

where $\gamma(.)$ is a discrepancy function. $\gamma(.)$ is equal to zero if $x_{ak}x_{yk}x_{bk}$ are the same symbols or represent a monotonic transition. $\gamma(.)$ increases with the minimum distance of x_{yk} w.r.t. the set of symbols representing a monotonic transition between x_{ak} and x_{bk}.

4.2. Multiple decoders

A multiple decoder scheme as shown in Figure 1 will also be considered. Each decoder uses models of different units attempting to capture types of regularities in phoneme strings, phonotactics, environment knowledge. Different acoustic parameters and recognition paradigms (HMMs, NNs, SVM) can also be considered, as well as, different acoustic features with, for example, variable time-frequency resolutions.

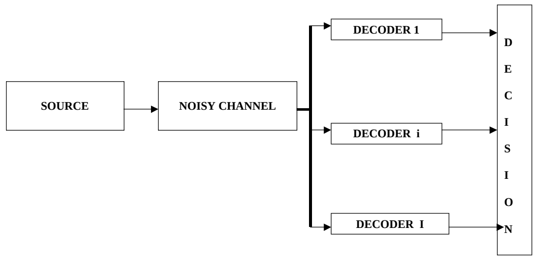

Fig. 1: Multiple decoder scheme.

Each decoder has a different set of phoneme models {syllable-based, pseudo-syllable based, diphone based, bigram, etc.} and generates K lattices of phoneme hypotheses: $P_k(A|f_i)$ and different strings of unit symbols. Rescoring of the word hypotheses for which the global score from the K decoders is maximum, could be either based on one lattice or based on a combination of the K lattices.

Acknowledgement

The work described in this paper is part of research effort carried out in the SMADA project on Telephone Directory Assistance [1]. This project is partially funded by the European Commission, under the Action Line Human Technology in the 5th Framework IST Programme.

The authors wish to express their thanks to Alexandre Ferrieux and Denis Jouvet from France Telecom R&D for the use of data.

5. REFERENCES

[1] Boves L., Jouvet D., Sienel J., de Mori R., Béchet F., Fissore L., Laface P. (2000). ASR for Automatic Directory Assistance: the SMADA Project. Proc. ASR, Paris, France.

[2] Byrne W., Knodt E., Khudanpur S. , and Bernstein J. (1998). Is Automatic Speech Recognition Ready for Non-Native Speakers? Proc. STiLL, pp. 37-40, Marholmen, Sweeden.

[3] Chomsky N., Halle M. (1968). The Sound Pattern of English. MIT Press.

[4] Cremelie N. and Martens J.P. (1999). In Search of Better Pronunciation Models for Speech Recognition. Speech Communication, 29(2-4): 115-136.

[5] Deligne S., Maison B., and Gopinath R. (2001). Automatic Generation and Selection of Multiple Pronunciations for Dynamic Vocabularies.. Proc. ICASSP. Salt Lake City, USA.

[6] Deshmukh N., Weber M., and Picone J. (1996). Automatic Generation of N-Best Pronunciations of Proper Nouns. Proc. ICASSP, Atlanta, US. pp. 283-286.

[7] Fosler-Lussier E. and Morgan N. (1999). Effects of Speaking-Rate and Word Frequency on Pronunciations in Conversational Speech. Speech Communication, 29(2-4):137-158.

[8] Fosler-Lussier E. (1999). Contextual Word and Syllable Pronunciation Models. Proc. ASRU Workshop, Keystone, USA

[9] Gao Y., Ramabhadran B., Chen J., Erdoğan H., and Picheny M. (2001). Innovative Approaches for Large Vocabulary Name Recognition. Proc. ICASSP, Salt Lake City, USA

[10] Jurafsky D., Ward, W. Jianping Z., Herold K., Xiuyang Y., and Sen Z. (2001). What Kind of Pronunciation Variation is Hard for Triphone to Model? Proc. ICASSP, Salt Lake City, USA.

[11] Kim T., Kang S., and Ko, H. (2000). An Effective Acoustic Modeling of Names Based on Model Induction. Proc. ICASSP, Istanbul, Turkey.

[12] Kim, J.H. and Woodland, P.C. (2000). A Rule-Based Named Entity Recognition System for Speech Input. Proc. ICSLP, Beijing, Republic of China.

[13] Korkmazskiy F. and Lee C.H. (1999). Pronunciation Modeling with Genetic Algorithms. Proc. ASRU Workshop, Keystone, USA. pp 528-531.

[14] Ramabhadran C., Bahl R.L., deSouza P.V., and Pandmanabhan, M. (1998). Acoustic Only Based Automatic Phonetic Baseform Generation. Proc. ICASSP, Seattle, USA.

[15] Riley M., Byrne W., Finke M., Khudanpur S., Ljolje A., McDonough J., Nock H., Saraclar M., Wooters C. and Zavaliagkos G. (1999). Stochastic Pronunciation Modeling from Hand Labeled Phonetic Corpora. Speech Communication, 29(2-4):209-224

[16] Suchato A. (2000). Framework for Joint Recognition of Pronounced and Spelled Proper Names. MS Thesis, MIT. http://www.sls.lcs.mit.edu/sls/publications/index.html

[17] Strik H. and Cucchiarini C. (1999). Modeling Pronunciation Variation for ASR: A survey of the literature. Speech Communication, 29(2-4):225-246.

[18] Witt S. and Young S. (1999). Off-line Acoustic Modelling of Non-Native Accents. Proc. Eurospeech, Budapest, Hungary. pp 1367-1370.