DYNAMIC GENERATION OF PROPER NAME PRONUNCIATIONS FOR DIRECTORY ASSISTANCE

Frédéric Béchet, Renato de Mori, Gérard Subsol

LIA - University of Avignon, France

ABSTRACT

This paper deals with the difficult task of recognition of a large vocabulary of proper names in a directory assistance application. Rather than augmenting the lexicon with alternate pronunciations, which is unsuitable for very large vocabularies of proper names, a class of distortions of the canonical form is used as knowledge source (KS) for a new evaluation of the N- best hypotheses generated in a first recognition phase, in which new probability distributions are used. The KS is the result of applying constraints inspired by speech science to distortions obtained by automatic learning. Experiments on a very large French directory document the validity of the approach.

1. INTRODUCTION

Recognition of a large vocabulary of proper names in a Directory Assistant (DA) application is a difficult task of a very high perplexity. Furthermore, the phone sequences representing words are often imprecise because they are derived by text-to- speech programs and many names, especially foreign names, are represented by sequences which rarely correspond to what people say. This suggests to perform recognition as a multi- phase process in which the first phase generates hypotheses using available phonetic representations and plausible deviations from them are considered in successive phases.

A considerable amount of research has been performed in recent years on lexical modeling [1, 2, 3, 4, 5, 6, 7, 8]. The approach proposed in this paper does not require the availability of alternate pronunciations, but generates and scores plausible distortions during a progressive search. Among the previously proposed approaches which do not rely on a precompiled lexicon of alternate pronunciations it is worth mentioning the use of *dynamic lexicons*. A possibility to derive them is to train decision trees with different transcription probabilities for different word contexts [9]. The lexical representation used for decoding is dynamically determined based on the context. Details can be found in [9] 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 base form phones. Other form of context, such as word identity, speaking rate [10], 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 [11], 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 It has also been suggested that phoneme substitution is often taken into account in the mixture of Gaussians used in context-dependent phone models.

It has recently been pointed out that 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 [12]. Phoneme insertions, with high evidence in the acoustic data, are also worth considering.

The novel approach proposed in this paper is that possible distortion are dynamically generated as a refinement of representations available in a previous phase. Modifications are grouped into classes and, for each class, a search process uses specific class knowledge and probability distributions. Such a knowledge can be inspired by speech science.

Several rescoring methods, such as proposed in [13], have been proposed for DA applications. This paper describes a progressive search in which the N-best list of hypothesis, generated in a first search phase, is rescored by successively applying a knowledge source that systematically considers the possibility of inserting one phone into the canonical forms. New probability distributions are proposed, one based on the plausibility (in terms of speech science) of the insertion in a given context and the other based on the expected success of the insertion, given its acoustic evidence. Experiments are described that use the N-best hypotheses generated by a first step recognition, performed after a short dialogue with the user, by a system developed at France Telecom R&D for a very large French directory.

After the presentation of some theoretical considerations in Section 2, corpora and experimental results are described in Section 3.

2. ALTERNATIVE PRONUNCIATIONS AND RECOGNIZER ERRORS

Let W be the orthographic representation of a proper name hypothesized in a previous search phase and τ be a sequence of phonemes which is considered in the actual search phase as a possible modification of available representations.

In the experiments described in this paper, the available representation is a sort of canonical form of W and τ can be every sequence of phone segments obtained by allowing an insertion on the previous form. The following posterior probability is used for ranking name hypotheses:

$$P(W|A) = \sum_{\tau} P(W\tau|A) \approx \max_{\tau} P(W\tau|A) \qquad (1)$$

This project is partially funded by the European Commission, under the Action Line Human Technology in the 5th Framework IST Program

This implies, among other things, that, for a given τ, the path corresponding to the best alignment of the phone sequence model with the acoustic description A has to be found. In practice, this is not performed in the first search phase because non-admissible algorithms have to be used due to the very large vocabulary size.

On the contrary, the best alignment can be performed once the N-best candidates have been hypothesized and the scores for every phoneme at each frame are available. The posterior probability in 1 can be further expressed as follows:

$$P(W\tau|A) = \frac{P(A|\tau)P(\tau)P(W|\tau A)}{P(A)} \quad (2)$$

The recognized word can be obtained by the following decision rule:

$$W^* = arg \max_{W,\tau} \{P(A|\tau)P(\tau)P(W|\tau A)\} \quad (3)$$

The first term in equation 3 can be computed by multiplying the frame prior probabilities for every possible alignment of τ; $P(\tau)$ is the probability of a plausible distortion of the canonical form for W; while the computation of the last factor in the 3 is more difficult. Notice that, if distortions of the canonical form have to be considered, rather than generating a word model for each distortion, in the approach proposed here the search space is a matrix with the acoustic probabilities of each phoneme at each frame. In this way, insertions, deletions, substitutions, and different segmentations for each form can be evaluated with the same search algorithm because different forms just influence the search constraints.

The following sections present a method for selecting acceptable distortions τ according to both knowledge based constraints and automatic learning techniques on a development corpus.

2.1. Choosing acceptable distortions

Choosing a suitable τ to be used in the search process is crucial, since not all the possible sequences can be considered. Knowledge-based constraints are useful for selecting plausible distortions according to some general simple rules. For example, it is well known, in French, that silence $[sil]$ segments can be inserted between certain pairs of phonemes (this is frequent for certain pairs of consonants, like $[m][n]$) and that a schwa can be inserted after a consonant at the end of a utterance or between an occlusive and a nasal consonant.

Some further considerations are useful for this purpose. Let η_W be the canonical form for word W. Let δ_W be the distortion corresponding to the optimal insertion of a phone for W. It is possible to write:

$$P(\tau) = P(\eta_W \delta_W) = P(\delta_W|\eta_W)P(\eta_W) \quad (4)$$

A rough estimation of $P(\delta_W|\eta_W)$ can be derived from the literature, by assigning a high value to insertions that have been indicated as plausible ones and a low value (even zero) to those that have never been hypothesized or appear to be impossible based on speech production knowledge. Useful suggestions are found in important speech science papers and particularly in [14].

At the moment, as reliable probability estimations are not available, a uniform, non zero probability has been assumed for plausible insertions and a null probability has been assumed for the other cases. $P(\eta_W)$, the probability of the canonical form of a name has been assumed to be uniform as reliable statistics of name use are not yet available. A suitable fudge factor should also be used for $P(\tau)$.

2.2. Evaluation of a given distortion

Various approximations can be considered for computing the probability corresponding to the last factor in equation 3. In a distortion τ, if a phoneme $[y]$ is inserted between string B and string E, two cases are of interest, namely when the insertion was successful, represented by the binary predicate $s(W|ByE, A)$ and when the insertion was unsuccessful, represented by the complement of predicate s.

A reasonable assumption, in case of insertion, is the following:

$$P(W|\tau A) = P[s(W|ByE, A)] \quad (5)$$

The computation is still difficult, because A has a very large variability, which suggests grouping different instances of A into classes. The simplest solution consists in having only one class. In such a case, one gets:

$$P[s(W|ByE)] = \frac{c[s(W|ByE)]}{c[s(W|ByE)] + c[\overline{s}(W|ByE)]} \quad (6)$$

where c indicates the counting function and $\overline{s}$ indicates unsuccessful insertion. An analogous probability has to be computed for the case in which the successful form is the canonical one. Notice that, even if A does not appear in equation 6, the probability of success for an insertion depends on A. Methods which allow to manipulate several classes of A are discussed in section 4.

The counts of equation 6 are estimated on a development corpus by using the following algorithm:

1. consider for τ all possible elements of classes of distortions (in the experiments reported here there is only the class of one phone insertion),
2. ignore class elements (e.g., possible insertions) which do not satisfy constraints derived from speech science knowledge,
3. select the best τ by searching for the best alignment (using acoustic models only) of all acceptable candidate sequences compatible with the constraints; if the likelihood of the best τ is not superior to the likelihood of the best transcription of the previous search phase (e.g., the canonical form), then replace τ with the best transcription of the previous search phase,
4. rescore the name candidates by using equation 3.

In this experiment, any insertion which leads to an improvement in the rescoring process of the N-best development corpus is considered successful. Similarly, any insertion which decrease the rank of the reference word after the rescoring process is considered unsuccessful.

3. EXPERIMENTAL RESULTS

3.1. Experimental setup

A development and a test corpora were used, each consisting 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.

Rescoring is based on a A^* decoding strategy. The search space is represented by a matrix of the acoustic model probabilities of each phoneme in each frame. These probabilities are derived from a lattice of phoneme hypotheses generated in the first phase. This type of rescoring benefits from the possibility of performing an admissible exhaustive search in a very short time because acoustic models are no longer used.

An experiment was carried out by considering only the following types of insertions, called *Knowledge Source* (KS) insertions:

- a silence $[sil]$ between pairs of phonemes, except when the second phoneme is vocalic,
- a schwa after a consonant at the end of a utterance,
- a burst recognized as a weak fricative or an unvoiced stop at the end of a utterance.

As very few data were available for estimating the model probabilities, they were initially set to 1 only for the insertions compatible with the KS and zero otherwise. The results for the development set using only this KS are shown in the second column of table 1, while the results obtained in the first phase are shown in the first column. The correctness percentage of rank i corresponds to the occurrence of the correct name in the first i elements of the 5-best.

N-best set	1^{st} pass	KS insertion
1	500 (50.0%)	598 (59.8%)
2	+110 (61.0%)	+62 (66.0%)
3	+39 (64.9%)	+19 (67.9%)
4	+29 (67.8%)	+9 (68.8%)
5	+14 (69.2%)	+4 (69.2%)

Table 1. Performance (% correct) obtained by rescoring proper names considering KS phoneme insertions

The improvement due to rescoring comes from a more accurate segmentation performed with an almost admissible search algorithm. Nevertheless, a significant portion of it results from rescoring with insertion assumptions even with simple insertion types and uniformity assumption for some probability distributions.

In a second experiment, expression 6 was then applied to the development set as presented in section 2.2. The insertions considered are the ones that, when applied only to the correct answer, allowed it to move to the first position when it was below it. These insertions were then applied to all the development set and to all the candidates in it allowing also errors to jump to the first position when they were below it. The results are shown in table 2, column 3.

The same gain of about 10% absolute, between the first and second pass, can be observed between the development and the test corpus. The training of the insertion model by mean of a development corpus produces a large gain in precision on this development corpus and, even if the gain is smaller, the same trend is observed on the test corpus.

	1^{st} pass	KS insert.	succ. insert.
1^{st} rank devt. set	500	598	630
1^{st} rank test set	480	588	595

Table 2. Performance (% correct) obtained by using a development corpus for selecting insertions

4. PERSPECTIVES

The results presented in the previous section show that, if the KS distortions seem very robust, not all of those obtained on the development corpus can be successfully applied on the test corpus (the absolute gain of 3.2% on the development corpus dropped to less than 1% on the test corpus). We believe that this lack of robustness is due to the simplification done in equation 6, by considering only one class for the acoustic parameter A.

Nevertheless, a limited number of classes for A can be obtained by describing the behavior of phonetic feature probabilities in the time segment of the inserted phoneme. This paragraph describes an ongoing work on using such phonetic features in order to increase the robustness of the insertion selection algorithm.

A set of binary phonetic features proposed in [15] for the French language has been used.

Each time frame can be represented by a vector of feature posterior probabilities : $X_t = [x_{1t}, .., x_{kt}, .., x_{Kt}]$

The posterior probability for the k^{th} feature in the t^{th} frame is given by:

$$x_{kt} = \frac{\sum_{f \in T_k} P(A_t|f)P(f)}{\sum_{f \in T_{all}} P(A_t|f)P(f)} \quad (7)$$

T_k is the set of phonemes for which the k^{th} feature has value true, T_{all} is the set containing all the phonemes.

The analysis of the time evolutions of feature posterior probabilities allows one to consider two classes for A (once again, we take as example the situation of a phoneme $[y]$ being inserted between string B and E):

- the first class, that will be indicated by N, contains cases in which the time evolutions clearly show feature changes which are natural transitions between B and E and justify the hypothesization of $[y]$ just because not all the feature probabilities change at the same time;
- the second class, indicated as $\overline{N}$, represents the fact that some probabilities exhibit a time evolution which denote the intention of uttering $[y]$ and deviate from the transition between B and E.

For example, this happens if a feature probably rises because the feature value is true for $[y]$, while it is constantly false in the last phoneme of B and in the first phoneme of E.

Using feature evidence, two classes can be used for the computation of equation 5:

$$P[s(W|ByE, A)] = P[s(W|ByE, \overline{N})]P(\overline{N}|A) + P[s(W|ByE, N)]P(N|A) \quad (8)$$

$P[s(W|ByE,\overline{N})]$ and $P[s(W|ByE,N)]$ can be computed by counts as for equation 6, while $P(N|A)$ can be expressed in terms of evidence of peaks and valleys of feature posterior probabilities in the segment of $[y]$.

Another possibility is that of extracting from the signal a set F of acoustic cues which are suitable for separating successful and unsuccessful insertions. An interesting approach for feature extraction using neural networks has been recently proposed in [16].

In such a case, indicating with $y(B,E)$ the insertion of y between B and E, $P_S[y(B,E)]$ a successful insertion and $P_{\overline{S}}[y(B,E)]$ an unsuccessful insertion, the computation of equation 5 can be carried out as follows:

$$\begin{aligned} P(W|\tau,A) &= \frac{P_S[y(B,E),F]}{\sum_{V\in lexicon} P(V,\tau,F)} \\ &\approx \frac{P_S[y(B,E),F]}{P_S[y(B,E),F]+P_{\overline{S}}[y(B,E),F]} \\ &= \frac{P_S[F|y(B,E)]}{P_S[F|y(B,E)]+P_{\overline{S}}[F|y(B,E)]\times\rho} \end{aligned} \quad (9)$$

with

$$\rho = \frac{P_{\overline{S}}[y(B,E)]}{P_S[y(B,E)]} \quad (10)$$

Successes and failures can be assumed to have Gaussian distributions that can be inferred from experiments, while ρ can be estimated with counts of successes and failures.

5. CONCLUSIONS

Proper name recognition for DA applications is approached in this paper by progressive search in which the N-best list of hypothesis, generated in a first search phase, is rescored by successively applying knowledge sources. The first knowledge source propose in this paper, systematically considers the possibility of inserting one phone into the canonical forms.

New probability distributions are proposed, one based on the plausibility (in terms of speech science) of the insertion in a given context and the other based on the expected success of the insertion, given its acoustic evidence. Preliminary experimental results show the effectiveness of this search phase.

6. ACKNOWLEDGEMENT

The work described in this paper is part of research effort carried out in the SMADA project on Telephone Directory Assistance [17]. This project is partially funded by the European Commission, under the Action Line Human Technology in the 5th Framework IST Program.

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

7. REFERENCES

[1] D. Neeraj, M. Weber, and J. Picone, "Automated generation of n-best pronunciations of proper nouns," in *ICSLP'96, Seattle*, 1996.

[2] Cremelie N. and Martens J.P., "In search of better pronunciation models for speech recognition," in *Speech Communication, 29(2-4): 115-136.*, 1999.

[3] Gao Y., Ramabhadran B., Chen J., Erdogan H., , and Picheny M., "Innovative approaches for large vocabulary name recognition," in *ICASSP2001, Salt Lake City, USA*, 2001.

[4] J.H. Kim and P.C. Woodland, "A rule-based named entity recognition system for speech input.," in *Proc. ICSLP, Beijing, Republic of China*, 2000.

[5] Kim T., Kang S., , and H. Ko, "An effective acoustic modeling of names based on model induction," in *ICASSP2000, Istanbul, Turkey*, 2000.

[6] Strik H. and Cucchiarini C., "Modeling pronunciation variation for asr: A survey of the literature," in *Speech Communication, 29(2-4):225-246*, 1999.

[7] Ramabhadran C., Bahl R.L., deSouza P.V., , and M. Pandmanabhan, "Acoustic only based automatic phonetic baseform generation," in *Proc. ICASSP, Seattle, USA*, 1998.

[8] Riley M., Byrne W., Finke M., Khudanpur S., Ljolje A., McDonough J., Nock H., Saraclar M., Wooters C., and Zavaliagkos G., "Stochastic pronunciation modeling from hand labeled phonetic corpora," in *Speech Communication, 29(2-4):209-224*, 1999.

[9] E. Fosler-Lussier, "Contextual word and syllable pronunciation models," in *Proceedings of the 1999 IEEE ASRU Workshop, Keystone, Colorado*, 1999.

[10] E. Fosler-Lussier and Morgan N., "Effects of speaking-rate and word frequency on pronunciations in conversational speech," in *Speech Communication, 29(2-4):137-158*, 1999.

[11] S. Deligne, B. Maison, and R. Gopinath, "Automatic generation and selection of multiple pronunciations for dynamic vocabularies," in *ICASSP 2001, Salt Lake City, UT*, 2001.

[12] Jurafsky D., W. Jianping Z. Ward, Herold K., Xiuyang Y., , and Sen Z., "What kind of pronunciation variation is hard for triphone to model?," in *ICASSP2001, Salt Lake City, USA*, 2001.

[13] Bouwman G. and Boves L., "Using discriminative principles for recognising city names," in *Proc. Eurospeech, Aalborg, Denmark*, 2001.

[14] J. Ohala, "Sound change as nature's speech perception experiment," in *Speech Communication, 13:155–161.*, 1993.

[15] F. Dell, "Les regles et les sons," in *Paris: Hermann*, 1985.

[16] Lee K.T. and Wellekens C., "Dynamic lexicon using phonetic features," in *Proc. Eurospeech, Aalborg, Denmark*, 2001.

[17] L. Boves, D. Jouvet, J. Sienel, R. de Mori, F. Bechet, L. Fissore, and P. Laface, "Asr for automatic directory assistance: the smada project," in *Proc. ASR2000*, 2000.