Issues in speech recognition applied to directory listing retrieval

Isabel Trancoso*, Carlos Ribeiro, Ricardo Rodrigues*, Miguel Rosa***

* INESC/IST **INESC/ISEL
INESC, R. Alves Redol, 9, 1000 Lisbon, Portugal E-mail: imt@inesc.pt

Abstract

This paper addresses several issues relevant to the application of speech recognition in directory listing retrieval: the very large dimension of the vocabularity, the confusability between vocabulary words and the powerful syntactic models implicit in full names. These issues will be addressed using as a case study the automation of the directory assistance of the two largest cities in Portugal.

1. INTRODUCTION

Directory listing retrieval services are currently handled by hundreds of human operators around the world. Their full automation requires both speaker-independent recognition and text-to-speech capabilities. Our goal in this paper is restricted to the first technique and most particularly to the characteristics of this type of service which distinguish it from other current automated operator services. In the past few years, speech recognition has advanced considerably, allowing the implementation of a number of small vocabulary, isolated-word ASR applications through the telephone network, with word-spotting and barge-in capabilities, increased robustness to environmental noise and channel errors, etc. There are several issues, however, which make the automation of directory listing retrieval different in terms of speech recognition: the dimension and nature of the vocabulary, the syntax model and the type of dialogue required. The paper will address these specific problems, using as a case study the automation of the directory assistance services of the two largest cities in Portugal: Lisbon and Oporto.

The most obvious problem is the very large dimension of the vocabulary, which implies the use of phoneme-based statistical models, instead of whole-word models. The vocabulary itself poses specific problems: on one hand, the pronunciation of proper names may imply the use of pronunciation rules different from the ones generally adopted for the common lexicon. On the other hand, the vocabulary includes highly confusable proper names which render human recognition difficult.

The study of the pronunciation of proper names was the goal of the LRE Onomastica project, now in its ending phase. Most of the pronunciation problems encountered in the national database concerned foreign origin names and company names formed by acronyms [7]. In fact, names of foreign origin may have significantly different "nativised" pronunciations. Some of them have entered the lexicon a long time ago and have since then undergone different degrees of adjustment to the language sound structure. For others, however, the pronunciation depends largely on the speaker's familiarity with the original language and a number of other complex factors. The company names formed by acronyms also constitute a problem for our language, as they may follow pronunciation rules which significantly differ from the ones observed for the common lexicon and may show a great deal of variability in pronunciation among native speakers.

The study of the confusability of proper names was partly motivated by our participation in the collection of a large telephone database of the "Polyphone" type, together with one of the national Telecom operators (European project SPEECHDAT). Although not mandatorily included in the project, a subset of this generic database will include read and spelt names of persons and also spontaneous answers simulating a directory enquiry. The design of this part of the database motivated a study of the confusability of proper names.

The difficulties arising from the dimension and the nature of the recognition vocabulary in automated directory retrieval are to a great extent counterbalanced by the powerful syntax models provided by full names. This is illustrated by the 94% name accuracy reported in [2] for a speaker-independent enquiry system with $18,000$ names operating over a dialed-up telephone line, using isolated letters for spelling the name. Despite the high confusability of similar letters, the system could handle common misspellings of names with insertions or deletions of a single letter, two letters in reverse order in a given string or single letter substitutions.

The paper will start by a description of our case study, including the types of requests that must be contemplated, some relevant characteristics of the database, and also the latest progress in the study of the pronunciation of foreign names and acronyms in our language made in the Onomastica project. It will then address the word confusability problem for proper names and, before concluding, it will also briefly report on our first attempt at a fully automated directory assistance service which was based on touch-tone technology combined with the prerecording of the most frequent proper names in the directory [4].

2. Case Study

The directory assistance services of Lisbon and Oporto currently occupy around 150 operators, distributed through several operator centers, who attend almost 50,000 enquiry requests daily. On average, 25% of these requests are reverse telephone directory enquiries, i.e., the client indicates the telephone number of the person whose name/address he/she wishes to find. The remaining 75% are direct enquiries statistically distributed as follows: 30% indicate the name(s) of the person whose phone number/address is desired; 20% indicate name and street and wish to obtain the telephone number; 20% indicate name and full address (including the numbers of the door and floor) and the remaining 5% indicate besides name and address also some additional pieces of information like profession (Dr., Engineer), name of building, type of subscriber (individual/company), etc.

The database used in these services includes names of persons, streets, towns and companies of the two areas, amounting to a total of around 100,000 different isolated words. This corpus was processed in the context of the Onomastica project, by eliminating many orthographical errors and appending to each entry, its frequency of occurrence, its etymology, its category (first name, surname, street, company, town or region, common word, company name and acronym), and also its (broad) phonetic transcription(s). Multiple categories are also allowed. In the original list of 100,000 words, roughly 50% are unique occurrences. Only 13,000 words occur more than 10 times and only 2,700 occur more than 100 times. Using the first of these subsets, a coverage of about 84% of the full names in the telephone directories of the whole country can be obtained.

Portugal has approximately 10 million inhabitants among which a relatively small percentage have foreign names. In fact, not only there are few immigrants but also until very recently the nativisation of the orthography of foreign names (e.g. Katya - Catia) was forbidden by law. In order to distinguish people by name, each person has typically one or two first names and two or more last names, at least one from each parent side. In the Lisbon directory, for instance, only 7% of the persons/firms have less than 3 names, 34% have 3 names, 36% have 4 names, 16% have 5 names and 7% have more than 5. Liaison particles are also relatively frequent, although they have not been included in the above percentages: 2% of the names have the conjunction *e* ("and") and 5% have the preposition *de* ("of") or its contractions (*do*, *da*, *dos*, *das*). These liaison particles impose the adoption of sandhi rules unless an isolated word dialogue is selected.

An interesting characteristic of proper names in Portuguese which distinguishes them from the common lexicon is the relative high percentage of certain word endings. In fact, among the most frequent 13,000 names of persons, around 11% end in *eiro* (or its gender and plural forms *eira*, *eiros*, *eiras*) or *inho*, a diminutive suffix with gender and plural forms included as well. Other names ending in *indo*, *ino*, *ano*, *ílio*, *ário* have also frequently a female correspondent ending in *a* instead of *o*. These common word endings are all stressed in the syllable before the last and significantly increase the confusability between proper names. It is also interesting to notice that, for the above mentioned list, 11% of the words have both gender forms in the list, and 6% have plural and singular forms.

Among all the classes of proper names in the directory, company names formed by acronyms are one of the most troublesome in terms of speech recognition and synthesis. Acronyms constitute 38% of the most frequent 50,000 entries of the original list of 100, 000 different words. Besides showing some differences relative to the common lexicon in terms of orthography, their main difficulty concerns the large variability that can be found in their pronunciation among native speakers. Most of this variability is related to the phonetic realisation of unstressed vowels. In Portuguese, vowel raising is usually applied to vowels in pre-stressed positions. Many acronyms, however, are formed by compounding, a lexical process which is not too common in our language and which is not, therefore, easily recognizable. As speakers become more conscious of this type of formation process in acronyms, they tend to analyse every form as compound, assigning stress to each element that may coincide with a root or word or which may be interpreted as a truncation of either of these. Since stressed vowels do not undergo reduction, sequences of syllables with open vowels became very frequently. This may explain the appearance of a general strategy of not allowing vowel raising in pre-stressed positions. This strategy is systematically adopted in all cases which have endings that characterize this class of names. In the remaining cases, however, the use of vowel raising or not varies too much, thus imposing the adoption of a relatively large number of possible phonetic transcriptions. Because of these difficulties, company names were temporarily excluded from the following confusability study.

3. Word confusability

There are several models for predicting confusability [5], [3], [6] whose comparison for Portuguese would be very interesting. Due to the time constraints of the project, however, we have tried to derive a simplified word-confusion model, based on the similarity between the phonetic transcriptions of word pairs, which takes into account the phone confusion matrix measured by a phone-based recognizer. This simplified model is an adaptation of one of the classic similarity measures between pairs of character strings, typically used in the automatic detection and correction of errors in written words [1]. Many of the existent methods for determining the similarity between two words can be described in terms of operations on a coincidence matrix A, i.e., a matrix whose element a_{ij} is either a one or a zero, depending upon whether the ith element of the first word W_1 matches or not the jth element of the second word W_2. These operations can be decomposed into three stages: in the first one, a weighting function is applied to all the matrix elements, according to some estimate of the likelyhood of their representing a true relationship between two elements of a pair; in the second (optional) stage, a selection procedure is applied to provide a matrix with only one non-zero element in each row and column; finally, a similarity function is computed from the resulting matrix. Among the multiple combinations of techniques for the different stages, we chose the following one:

- Weighting function - the selected function replaces each element of the coincidence matrix by the product of that element and one minus the distance between that element and the axis (defined as the line through the points $(1,1)$ and (m,n) for an mxn matrix), i.e., the weights are given by:
  $w_{ij} = a_{ij} - (1 - \mid (i-1)/(m-1) - (j-1)/(n-1) \mid)$
- Selection function - this function is applied to the weighted coincidence matrix, by selecting the largest element of the matrix, then the largest element in the remaining rows and columns, etc.
- Similarity function - this final function computes a cumulative sum S over sets of diagonally consecutive elements (i.e., any subset of elements whose difference of indices $i-j$ equals some constant k). In any such set of n elements, the one with lowest indices is added n times, the second $n-1$ times and so forth, the nth element being added just once. The distance is computed as
  $d(W_1, W_2) = 1 - S/((l^2 + l)/2)$
  where $l = max(m, n)$, and the normalizing factor equals the sum S for an identity matrix, thus resulting in a null distance, as required for two identical words.

This 3-step procedure was adapted in the following way: first, the phonetic transcriptions corresponding to the two words are retrieved from a table which was previously automatically generated by a grapheme-to-phoneme conversion module and later manually corrected. Assuming, for the time being, that each word has a single transcription, let us denote by T_1 and T_2 the corresponding strings. Next, a coincidence matrix was built, with each phonetic symbol playing the role of an element. However, instead of ones and zeroes, the matrix is build by replacing each element a_{ij} with the probability of element i being recognized as element j, computed using a HMM phone-based recognizer. The application of the 3-step procedure to this asymmetrical coincidence matrix obviously yields different results when computing the distance between word $W-1$ and word W_2 or vice-versa. The word confusability was thus defined as:

$Conf(W_1, W_2) = 1 - (d(T_1, T_2) + d(T_2, T_1))/2$

When each word has multiple transcriptions, the same procedure is applied to all possible ones, and the largest distance values are selected for the computation of the confusability score. Given the large computational effort required by the computation of $13,000_2$ pairs of names, the computation was restricted to the most frequent $1,000$ names. The confusability score is maximum for homophones, obviously. The class with highest scores, next to this one, includes male-female pairs of the same name, closely followed by singular-plural forms, as expected. By dividing the $[0,1]$ range of scores linearly by 10 classes, the class with the lowest confusion scores includes 87% of the possible pairs, the second lowest one includes 11%, and the remaining ones, only 2%.

Both the computation of the confusability measure itself and the comparison of these predicted results with substitution probabilities obtained in real recognition tests were strongly affected by the lack of large labelled corpora in our language. This work is still in progress, which motivated the use of a quasi-automatic procedure for computing the phone-confusion matrix. For the phones which were close to the English ones, initial models were derived from the labelled TIMIT corpus (New York City dialect). For the remaining ones, the manually labelled CVC subset of EUROM.1 was used. The training was done using Baum-Welch reestimation, and 3-state, 3-mixture-per-state models. Each input vector had 26 coefficients (12 cepstrum, 12 delta-cepstrum, energy and delta-energy). These initial models were then used to automatically align a subset of the Portuguese EUROM.1 corpus (the 150 passages read by the 10 subjects of the Few Talkers group), using the Viterbi algorithm. This automatically aligned corpus was then used to retrain 45 new phone-based models and realign the phonetic transcriptions, in a bootstrap procedure which was iterated a couple of times. Finally, triphone models were also trained. The training and testing was done using the HTK software.

The results of the simplified confusion score, using this phone-confusion matrix were compared with the probability of pair-wise confusions in a very preliminary test, using only 10 speakers. 40 pairs of words were selected between the 10 classes, in order to include a few from each class. The subjects were asked to read this list of 80 names, twice, in a quiet environment. Many recognition errors can only be attributed to badly trained models. However, when the word itself is excluded from the recognition vocabulary, the system often recognizes a word with small predicted distance from that one. This measure is obviously inferior to the one described in [6], whose application was also conditioned by the existence of a labelled database for Portuguese large enough to estimate phone confusions in their phonetic contexts. This work is still in progress, and will allow in the near future the computation of more meaningful confusability scores.

4. Directory retrieval based on DTMF technology

An experimental service for directory listing retrieval was developed by INESC to fully automatize the database query service developed for Portugal Telecom, which is currently being used by the human operators. The reverse service posed no particular problems. For the direct service, however, the use of the telephone keypad required the association of letters to each of the digits. In Portugal, no standard is defined for this association, which causes a relatively large variety of keypads ranging from the U.S. standard with only 24 letters sequentially distributed in groups of 3 among 8 of the digits, to the telephone keypads of several GSM manufacturers using the whole alphabet distributed through 9 digits. This motivated the study of the optimum letter-to-digit association for Portuguese which minimizes the ambiguity caused by the many-to-one association, for the above mentioned corpus of 13,000 most frequent isolated names. For the optimum association, the percentage of names which correspond to ambiguous digit sequences (more than one word) is only 4%. Disambiguation is done by using the powerful syntax models provided by full names. The system has been operating successfully for some months on an experimental basis at INESC, being able to detect the type of keypad used. The user is prompted for additional names without any particular order every time the system finds more than 5 subscribers with the same name(s). The relatively low penetration of DTMF and the difficulty of typing letters in a small keyboard are strong motivations for replacing the touch-tone technology with speech recognition as we aim to in this study.

5. Conclusions

This paper described some of the potential problems of the application of speech recognition to directory listing retrieval. It presented a simple measure of the confusability between word pairs and briefly described an experimental automated directory assistance service where DTMF technology replaces speech recognition, by assigning two or three letters to each of the keys of the telephone keypad. The powerful disambiguation capabilities of full names may play a very important role in the application of speech recognition to this type of services in the near future.

6. Acknowledgements

This paper addresses some issues which were previously dealt with during the Onomastica project and the development of the experimental directory service based on DTMF technology. For their help in these projects, the authors would therefore like to thank their colleagues Céu Viana, Luís Oliveira, Gonçalo Marques, Isabel Mascarenhas and Manuel Mira, and also Portugal Telecom for the raw data.

7. REFERENCES

[1] C. Alberga, *String Similarity and Misspellings*, Comm. ACM, 10(5), pp. 302-313.

[2] B. Aldefeld, L. Rabiner, A. Rosenberg, and J. Wilpon, *Automated Directory Listing Retrieval Based on Isolated Word Recognition*, Proc. IEEE, 68(11), pp. 1364-1379, November 1980.

[3] B. Lindberg, *Recogniser Response modelling from testing on series of minimal word pairs*, Proc. ICSLP'94, pp. 1275-1279.

[4] G. Marques, I. Trancoso, L. Oliveira and A. Serralheiro, *Automated telephone directory services (in Portuguese)*, Actas do 1.o Encontro Nacional do Colégio de Engenharia Electrotécnica, pp. 435-440.

[5] R. Moore, *Evaluating Speech Recognisers*, IEEE Trans. on ASSP, 25(2), April 1977.

[6] D. Roe and M. Riley, *Prediction of Word Confusabilities for Speech Recognition*, Proc. ICSLP'94, pp. 227-230.

[7] M. Céu Viana, I. Trancoso and F. Silva, *On the Pronunciation of proper names and acronyms in European Portuguese*, 2nd Onomastica Research Colloquium, Dec. 1994.