Toward Automatic Adaptationof the Acoustic Models and of the Formulation Variants in a Directory Assistance Application

Toward Automatic Adaptation of the Acoustic Models and of the Formulation Variants in a Directory Assistance Application

M. Andorno, P. Laface, C. Popovici Γ, *L. Fissore, C. Vair* K

Γ Politecnico di Torino, Italy
{andorno,laface}@polito.it
K Loquendo, Torino, Italy
(Cosmin.Popovici,Luciano.Fissore,Claudio.Vair}@loquendo.com

Abstract

The framework of this work is an already operational voice activated Directory Assistance (DA) service that allows a large amount of data from the field to be collected. Our goal is to improve its performance by adapting the acoustic models and the formulation variants of the system to the field.
For model adaptation, we propose to dynamically generate - without supervision - additional Hidden Markov Models tailored to the application environment and vocabulary. We report results showing significant improvements obtained in the recognition of the city names.
A relevant problem in DA for business listings is that customers formulate their requests for the same listing with a great variability. We show that an unsupervised approach allows to detect user formulations that were not foreseen by the designers, and that can be added, as variants, to the denominations already included in the system to reduce its failures.

1. Introduction

A nation wide automatic DA system is currently operational in Italy. This service, developed by Telecom Italia and Loquendo (formerly CSELT), routinely serves customers asking for both residential and business listings. Whenever the automatic system is unable to terminate the transaction with the customer, the call is routed to a human operator.
The first aim of this work was to find a procedure for generating and adapting to the application field, during the lifetime of the system, a set of models that could improve its performance, in particular for the recognition of city names.
It is well known that the integration of the information derived from real-life data with the knowledge embedded in the so called ``laboratory" acoustic models, typically trained with a large amount of read data, is a key factor to achieve high performance in application trials. Maximum A Posteriori estimation is one of the most successful approach to speaker and environment adaptation.
A large amount of user interactions can be easily collected by an operating DA system. This reduces the problems related to the lack of training data, rather, it raises the problem of how to filter a huge amount of speech data. Moreover, since it would be expensive to have a human operator label every utterance, the data collected from the field are labeled automatically. The resulting tokens are affected, thus, by labeling errors due to incorrectly recognized in-vocabulary or out of vocabulary words.
The Italian telephone book listings includes more than 25.000.000 records, about 3.500.000 of which are business listings. Since about 80% of the DA accesses is related to business listings, it is important to improve the percentage of success of the automatic DA system for this class of calls.
One of the hardest problems to be solved is that customers formulate their requests for the same business listing with great variability. In our approach, the requests for business listing are served by using a large vocabulary isolated word recognition technology, where the sequence of the words in a given entry is concatenated and transcribed as a single word, with possible silences in between. Since the content of the original records in the database does not, typically, match the linguistic expressions used by the callers, a complex processing step is needed for deriving a set of possible formulation variants (FVs) from each original record in the book listings.
The second aim of the paper is to present the evaluation of an approach towards automatic learning, from field data, of expressions typically used by customers to formulate their requests for the most frequent business listings.
We partition the field data referring to a given denomination into phonetically similar clusters from which new user formulations can be derived.
The paper is organized as follows: Section 2 gives a short overview of the Loquendo DA system. Section 3 reports how we select and train new models, and the obtained improvements. Section 4 recalls the steps for obtaining FVs from the records in the book listings and from field data. Section 5 details our approach for learning new formulations from field data, and discusses the obtained results. Finally, our conclusions are given in Section 6.

2. Loquendo DA system overview

The Loquendo DA application is based on large vocabulary isolated word recognition performed in two steps. The first step decodes the user utterance by means of a Hybrid HMM-NN model, where the emission probabilities of the HMM

This work was partially supported by the EU ITS Project SMADA Speech Driven Multi-modal Automatic Directory Assistance

states are estimated by a Multi Layer Perceptron. This step generates also the best phonetic string without lexical constraints. The second step, based on Continuous Density HMMs, decodes the same utterance using as a vocabulary the N-best hypotheses produced by the first step.
The combination of the hypothesis scores of the two steps, not only increases the recognition accuracy, but allows also the production of a reliability score for the best hypothesis. This score, and the phonetic one, are used by the dialog manager module for rejecting unreliable hypotheses or for reducing unnecessary request turns.
The system uses a set of vocabulary independent acoustic models that include the stationary parts of the context independent phonemes and all the admissible transitions between them. In particular, for the Italian language, we defined 27 stationary units and 348 transition units, for a total of 375 units. These models were trained using 26 hours of PSTN phonetically balanced read speech.
As a test-bed for the offline evaluation of the field adapted model we collected 8775 transcriptions of the turns related to the city name request. The system vocabulary includes 9325 city names.

3. Model selection and training

For the generation of the new models 38200 tokens were collected from the field and automatically transcribed using the laboratory models. Of these tokens, 19700 city names and 8800 province city names were selected according to their reliability score.
Unfortunately, the reliability score is not itself ``reliable" in the case of out of vocabulary words, thus part of field data still remains incorrectly labelled. The effect of these tokens on the model quality is, however, minimal as reported in [3].
Training of the field models is performed by an incremental MAP based approach using the laboratory models, or the previously computed models, for obtaining the necessary a priori knowledge.
It is worth noting that, in our approach, the new field dependent unit models are added to the original laboratory models, and that they are only used to transcribe the most frequent application words. Thus, the number of the system units slightly increases, and the most frequent words are transcribed in terms of two different set of units: a vocabulary independent set, and the other one application dependent.
If the occurrence of a word is much larger than the other ones, we activate a whole word training process. For these words the system will have yet another very specific model [3].
Further improvement of the recognition performance is obtained using features best suited to the telephone environment.

3.1. Single Step HMM

Table 1 shows the results of a single HMM recognition step. The first row reports the performance of the 375 vocabulary independent (VI) models introduced in Section 2. The units are modelled by left-to-right continuous density HMMs, with a maximum of 32 Gaussians density per state. A set of 380 triphones and 3 whole word models were added to the VI set to better cover the most important Italian cities. The features vector includes 39 parameters: 12 liftered and high-pass filtered MEL cepstral coefficients (MFCC), the log-energy and their first and second derivatives.

MODEL	WA %	Err. Reduction %
HMM-MFCC-12	81.9	REF
HMM-MFCC-VI	75.2	-36.7
HMM-MFCC-VD	80.7	-6.2
HMM-MFCC-VI+VD+WW	86.4	25.0
HMM-RLDA-VI	82.2	3.6
HMM-RLDA-VD	87.2	29.6
HMM-RLDA-VI+VD+WW	91.4	52.7

Table 1. Single step HMM recognition results

In absence of specific models (triphones and whole words) for the most frequent city names, the performance falls down as highlighted in the second row of the table. The exploitation of the data collected from the field to transform the 375 VI units into Vocabulary Dependent (VD) units improves the model quality as shown in the third row of the table.
The fourth row shows the results obtained by adding vocabulary dependent and whole word (WW) components to the VI models. In particular, we used 100 VD transcriptions and 20 WW models for the most requested city names. A significant error rate reduction is obtained as shown in the third column.
The last three rows report the results obtained using a feature vector of 36 parameters, derived from a Rasta PLP / LDA (RLDA) front-end. Comparing the figures with the corresponding results of the MFCC front-end a significant error rate reduction can be observed.

3.2. Single Step NN

The results of a single step Hybrid HMM-NN recognition with VI models are similar to the ones reported in the previous subsection, and are given in Table 2.

Models	WA %
NN-MFCC	81.9
NN-RPLP-VI	84.5

Table 2. Single step Hybrid HMM-NN recognition results.

3.3. Two steps NN+HMM

Table 3 summarizes the results of the two step strategy. During the first step a set of hypotheses is produced by the hybrid NN-HMM recogniser. In the second step the hypotheses are re-scored by means of the CDHMM system. The last column of the table includes figures about the reliability rate. The Correct Certainty (CC) is the percentage of samples correctly recognised and labelled as certain; the False Certainty (FC) is the percentage of samples incorrectly recognised and wrongly labelled as certain. Uncertain (UNC) gives the percentage of hypotheses labelled as not reliable.
The use of the improved models allows an overall reduction of the error rate of 25.3% and also an increase of the Correct Certainty (from 51.4 to 68.4). The increase of the CC is particularly important because it reduces the number of times the dialog manager is forced to ask the user for the province/region to identify the correct city name.

MODELS	% WA	% Error Reduction	%Reliability CC/FC/UNC
NN-MFCC + HMM-MFCC-12	87.2	REF	51.4/0.4/48.1
NN-RPLP-VI + HMM-RLDA-VI+VD+WW	90.4	25.3	68.4/0.5/31.1

Table 3. Two steps NN+HMM recognition results.

3.4. The effect of the training-set size on the recognition performance

To improve the quality of the vocabulary independent models, we extended the training sets to the 36 hours of data included in the SPEECHDAT databases.
As shown in Table 4, the inclusion of these data produces a 5.9% reduction of the error rate. In the case of gender dependent models the error rate reduction rises to 13.2%.
The forth row of Table 4 highlights the relevance of adding to the set of laboratory units a limited number of units trained from the field data: using only 7 hours of field data (versus 36 hours of SPEECHDAT) we obtained a 53.2% of error rate reduction.

HMMs	Training Hours	% WA	% Error reduction
RLDA	26	82.2%	REF
RLDA	26 + 36	83.0%	5.9%
RLDA Gender-Dep.	26 + 36	84.4%	13.2%
RLDA+VD+WW	26 + 7 Field	91.4%	53.2%

Table 4. Recognition results with more robust models

4. Generation of formulation variants

The generation of the formulation variants from the original book listing records is based on a semantic table that summarizes the content of the record, reducing the variability of the information inserted by the operators in the record fields.

Pref: 011 *Tel*: 5175296
City: TORINO *Prov*: TORINO
Address: 63, C. VITTORIO EMANUELE II
***Den*: BAR RISTORANTE LA FORCHETTA D'ORO DI MARIO ROSSI**
***Description*: PIZZA**
***Category*: RISTORANTI**

RISTORANTE LA FORCHETTA D'ORO
BAR LA FORCHETTA D'ORO
BAR RISTORANTE LA FORCHETTA D'ORO
PIZZERIA LA FORCHETTA D'ORO
LA FORCHETTA D'ORO

Figure 1*:* Example of some fields in a record, and its formulation variants

Using the semantic table information, a set of formulation variants with an associated score is produced. An example of record, and its generated formulation variants, ordered by score, is shown in Fig. 1. The best scoring formulation is also played back to the customer for confirmation.

Several turns for evaluating the coverage of the user formulations by the FVs were performed. In a first phase, real user data were collected from the interactions with human DA operators located in Turin. Then, other calls to a prototype DA, serving the Catania telephone district, were transcribed. From these preliminary tests it has been verified that the coverage of the original FVs was about 40%. It was thus mandatory to generate more accurate formulations for frequently requested listings, in particular for those presenting high failure rates.
Another large database (DB20000) was then collected from a month and a half of customer calls to an automatic system operating in Rome during the night. In particular, 8848 business calls, routed to the human operator by the system, because it was unable to deliver the desired information, were selected and transcribed. Another set was selected, and transcribed, from the daily traffic managed by 13 call centers distributed in several regions of Italy. All these calls correspond to the most frequently asked listings. The database includes a total of 20216 transcribed calls associated to the phone number provided by the human operators.
To generate new, more accurate, FVs, the transcribed denominations were analyzed, and generation rules derived, depending on the business category, according to a priori knowledge and data evidence. The FVs that received most attention were those related to hospitals, social services, public utilities, communication and transportation agencies, and the like, because they account for the majority of the calls. Since the automatic DA system is currently fully operational, new FVs, and possibly rules, are also derived whenever the service provider signals consistent anomalies.
By using the FVs rules derived from this new field data, the coverage of the FVs increased from 40% to more than 60%, using an average of 5 FVs per denomination. This also means that many users are rather collaborative and that the system prompt elicits concise linguistic expressions.

5. Automatic learning of formulation variants

5.1. Phonetic transcription

From the calls routed to the operators, the list of the most frequently requested phone numbers (provided by the operator) was selected. Setting a threshold of 20 requests per phone number, the most requested listings for the 3434 calls in the Catania database are 16 only. A much higher spreading has been experienced, as expected, for the nationwide calls, where 53 listings only were requested more than 20 times.
As said in Section 2, the recognition module of the system produces, together with the lexical constrained word hypotheses, the phonetic transcription of each utterance as the best sequence of phones obtained using a looped phone model. The phonetic strings associated to a given phone number are, thus, the automatic transcriptions of the different ways in which users formulate their request for a given business listing.
Table 5 shows, as an example, a small set of unconstrained phonetic transcriptions associated to the most requested phone number in the DB20000 database, corresponding to *FSInforma*, a widely used automatic train timetable information system, developed by CSELT, and managed by the Italian railways service provider Ferrovie dello Stato. These phonetic strings are widely different, and some of them can hardly be decoded. Recall, please, that these utterances were not com-

pleted by the automatic DA system for several reasons such as endpoint detection failures, extra-linguistic phenomena, low

ufiCoinfoRmaZionilstaZiuneditaeni	
enomaladelataZaneditoRenopoRtanovelomRaveRda	
feRoviodelostato	
ifuoRmaZionifeRoiedelostato	
esaZionetiboRtina	fveRuilstato
skaZione	oRaRiodetReni
tReno	fRovionalostato
nomaRomeRbefese	saZinoCentRale

Table 5: Samples of phonetic strings of user requests for the railway information service *FSInforma*

confidence scores, recognition errors due to the lack of a suitable transcription in the current database, etc. Another cause of system failures is that the user request was ambiguous, incomplete or embedded in a sentence, so that only after several turns of dialog with the user the operator was able to deliver the information.

On the other hand, in Table 5 it is possible to detect phonetic sequences that are easily interpreted since they are correct or nearly correct transcriptions of a denomination such as <feRoviodelostato> and <skaZione> for "Ferrovie Dello Stato" and "Stazione" respectively, and several variants with relatively few phonetic distortions.

It is also worth noting that, given a huge number of requests for the same phone number, there is a high probability of obtaining clusters of phonetically similar strings. The distance between two strings of phones can be obtained by Viterbi alignment of the two strings using the inverse of the log-probability of insertion, deletion and confusion among phones.

5.2. Clustering and selection of new formulations

For the most frequently requested phone numbers, each set of phonetic strings was clustered into similar subsets by using a furthest neighbor hierarchical cluster algorithm based on the mutual distance between each phonetic string.

The set of phonetically similar utterances is detected by stopping the bottom-up clustering when the number of elements that have been partitioned is greater than $N\text{-}N/(\log_2 N + 1)$, where N is the number of strings that are clustered. The clusters with few elements and large within cluster variance are discarded.

An interesting example of output of the clustering procedure is related to the main Catania Hospital, whose phone number is requested using several formulations referring to different clinics within the same hospital. Six denominations were added manually as formulation variants after the preliminary analysis of the errors of the prototype system in operation in Catania, The clustering algorithm detected the same formulations: a few elements of two clusters are shown in Table 6.

Ospedale Santa Marta	Ospedale Ferrarotto
<t&>ospedalafantamaRta	topelaleseRaRato
<t&>ospedalafantamaRta	sospelalefeRaRatoi
fospedalesantamasta	ostedavefeRanoto
	osbedaRefeRalato

Table 6: Two clusters for different formulations referring to the main Catania *Hospital*

Another example of automatic clustering, detecting a pronunciation variant of a foreign denomination, is given in Table 7, where some phonetic strings of two clusters related to requests for the French word *Auchan* are shown. As can be argued, the number of available samples for the Catania database is too small for deriving reliable phonetic transcriptions for new formulations. However, if a large enough database is available, it is possible to select significant clusters, characterized

French pronunciation with final schwa	Italian pronunciation
o&ana	au&an
vo&ana	au&an
fo&ana	auv&a
fo&an	aui&ian
o&a	au&aen

Table 7: Samples of two clusters of similar pronunciations for the denomination *Auchan*

Central element	Within - system nearest variant	No of elements	Cluster variance	Distance
feRoviedelostato	feRoviedelostato	156	2.13	0.00
staZioneCentRale	staZionefeRoviaRia	198	3.27	3.22
staZione	staZionefeRoviaRia	25	1.9	4.43

Table 8 – Central elements of the three significant clusters related to the denomination *FSInforma*

by high cardinality and small dispersion of the included phonetic strings. For example, using the 458 formulations that were available for the phone number of the *FSInforma* in the DB20000 database, the procedure generated several phonetically similar clusters. Only three of them, however, were significant according to a selection criterion related to the number of elements in the cluster (> 20 in this case) and to a low (< 4.0) dispersion of the elements within the cluster. The central element of the three clusters, defined as the string that has the minimum sum of the distance from all the other elements of the cluster, is shown in Table 8.

It is worth noting that, when the number of elements collapsed into a cluster is large enough, the central element of the cluster gives a very good transcription of the required denomination.

For the central elements in Table 8, good formulation variant candidates are the phonetic strings <staZioneCentRale> and <staZione> that are quite distant from the already present formulation <staZionefeRoviaRia>, while <feRoviedelostato> exactly matches a formulation already in the system.

6. Conclusions

We have shown that it is possible to improve the performance of a speech recognizer by dynamically adding to the set of laboratory units a small number of field trained models.

We have proposed an unsupervised approach for detecting user formulations that were not foreseen by the designers of a DA system. These formulations can be added to the system to reduce its failures.

7. References

[1] R. Billi, F. Canavesio, C. Rullent, Automation of Telecom Italia Directory Assistance Service: Field Trials results, Proc. IVTTA 1998, Turin, pp. 11-16, 1998.

[2] R. Gemello, D. Albesano, F. Mana, L. Moisa, "Multi-source Neural Networks for Speech Recognition: a Review of Recent Results", Proc. IJCNN-2000, Como, Italy, July 2000.

[3] C. Vair, L. Fissore, P. Laface, "Adaptation of Vocabulary Independent HMMs to an Application Environ-

ment”, Proc. of ICSLP, pp. II-839-842, Beijing, Oct. 2000.