USE OF SIMULATED DATA FOR ROBUST TELEPHONE SPEECH RECOGNITION

Coianiz Tarcisio, Falavigna Daniele, Gretter Roberto, Orlandi Marco
IRST - Istituto per la Ricerca Scientifica e Tecnologica
38050 Pante` di Povo, Trento, Italy.
Email: coianiz@itc.it, falavi@itc.it, gretter@itc.it, orlandi@itc.it
http://www.itc.it/irst/

6th European Conference on Speech Communication and Technology (EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ABSTRACT

The collection of telephone databases, for training speech recognisers, is a time consuming and costly work. In the paper we propose a method for producing simulated telephone data starting from clean wide band databases. The result of the simulation is the generation of a noisy database that can be used, in addition to other techniques, for compensating or adapting speech recogniser parameters with respect to different test environments. For the first of the two adopted test sets, performance improvements ranging from about 30% to about 9% have been measured, as a function of the quantity of real telephone data used, in addition to the simulated ones, for system training. For the second test set no significant improvements were obtained.

Keywords: simulated data, dialogue model, HMM

1. INTRODUCTION

The paper presents our recent work on the usage of simulated data for the retraining of Hidden Markov Model (HMM) of phone units. First, HMM parameters are bootstrapped from a band filtered version of APASCI [1], a clean speech database collected in our Labs. at 16 *kHz* sample frequency. Then, HMM retraining is carried out on increasing sets of speech files.

Initially the retraining list contains only non-telephone data, derived from APASCI itself after having properly processed each speech file in order to simulate both the telephone channel and noise. Then, two other different sets of telephone files are inserted in the retraining list and recognition performance is correspondingly evaluated. The objective of these experiments is to measure the effectiveness of adding a variable quantity of telephone data to the artificial ones for HMM retraining.

Obtained results show that a significant improvement in recognition performance can be achieved in the case of total lack of telephone data, that is when only simulated data are used for retraining. When more and more telephone data are inserted in the retraining list, the effectiveness of using simulated data correspondingly diminishes.

Tests have been led on two speech corpora collected during the usage of two systems developed in the past. One is an automatic switchboard [2], installed in an Office of *CARITRO* (a Bank in Trento), capable of recognising the names of all Bank employees (about 600) and branches (about 100). This system can connect the caller with the required person, or office, after his or her confirmation. The second system [2] is a prototype, used within our Labs., capable of handling a dialogue with a caller asking information about a train timetable. The dialogue strategy is of type *mixed initiative*.

The paper is organised as follows: Section 2 describes the training and test databases, Section 3 deals with the proposed simulation method, Section 4 gives some details on the dialogue architecture used for developing the previously mentioned train timetable prototype, Section 5 provides experimental results and conclusions are given in Section 6.

2. SPEECH DATABASES

2.1 Training databases

Speech signals contained in our telephone training databases were collected by an automatic system that called previously advised speakers. A typical call started with a short introduction. During this phase a background noise segment, having a duration of 500 *ms*, was acquired. Then, the speaker was asked to utter 4 different types of sentences: connected digits, city names, "*yes*" or "*no*" and phonetically balanced sentences. Recordings were carried out in two distinct phases, thus obtaining two databases: **PHONE1**, containing 3456 utterances from 190 speakers, and **PHONE2**, containing 2984 utterances from 269 speakers. Finally, a total of about 2000 short (500 *ms*) noise files resulted after the completion of the recordings.

As previously seen, for bootstrapping phone HMMs we use a band filtered version of APASCI (**AF**), which consists of 5215 phonetically balanced sentences. Then, retraining is carried out on the following increasing set of utterances:

- **APASCI-SIM** (**AS**), formed by 5215 utterances, derived by filtering the files of

APASCI with some measured telephone impulse responses and by successively adding a proper concatenation of short noise files (the details will be given in the next section);

- **APASCI-SIM** plus **PHONE1**;
- **APASCI-SIM** plus **PHONE1** plus **PHONE2**.

Recognition performance will be given for each one of the three training sets.

2.2 Test databases

Two different test sets were considered for the experiments. The first one (**CARITRO**) is formed by 1338 speech files. It was collected during the behaviour of the previously mentioned Bank internal switchboard and consists of utterances containing names of people, employed in the Bank, or names of branches of the Bank itself. The grammar (a Finite State Network) adopted for recognising this test is built by arcs in parallel, corresponding to all possible person or branch names, preceded and followed by optional networks representing several expressions for asking for people of branches. In this way, sentences such as: "*I want to speak to Mr. Smith*", "*Hallo, could you give me Mr. Smith?*", "*John Smith, please*", etc..., can be recognised.

The second test set (**FERROVIE**) is formed by 925 utterances, recorded during about 70 user interactions with the dialogue prototype, for train timetable inquiry, recently developed in our Labs [2]. In this case speech signals correspond to the various turns of the dialogues and are recognised with grammars specific of each turn. Performance for this test will be given in terms of semantic accuracy. In Section 4. the adopted dialogue model and grammars will be shortly described.

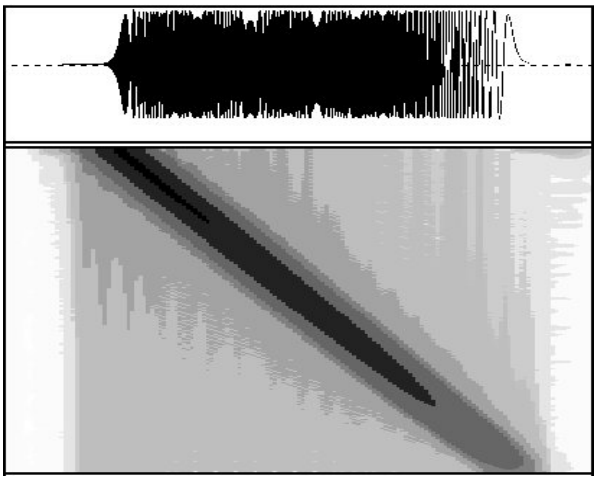
3. SIMULATION METHOD

The telephone noisy version of APASCI (**APASCI-SIM**) was produced by first filtering (and down-sampling) each wide-band clean file with a measured telephone impulse response and by successively summing a *noise* signal file.

Some telephone impulse responses were derived by means of a dual channel telephone board (RHETOREX). The two channels were connected to two different telephone numbers, linked together (one channel called the other) before doing the measures. A chirp signal (see Figure 1.), whose auto-correlation is an ideal impulse [3], was sent over one channel and was received and recorded from the other channel. The impulse response was obtained by the convolution of the recorded signal with the chirp [3][4].

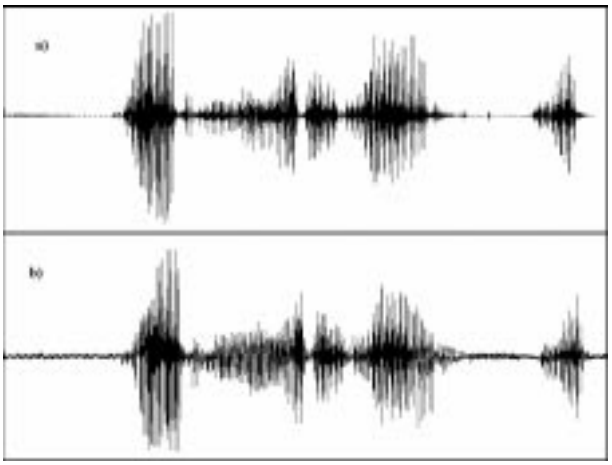
Each file of APASCI was filtered with one of the measured responses, then a *noise* signal was summed to the resulting filtered file. This latter *noise* was built by joining an appropriate number of the previously mentioned short noises, so that to cover the whole duration of the given file.

Figure 1.Chirp signal and corresponding spectrogram.



The amplitude of each short noise was properly scaled in order to obtain an uniform Signal to Noise Ratio (SNR) along the resulting noisy file. The SNR value was defined a priori for each file to process while, to reduce the spectral discontinuities, an Hamming window was applied to each short noise and the total *noise* to sum was derived by overlap and add. A portion of a speech signal of APASCI and the corresponding noisy version are shown in Figure 2.

Figure 2. Clean signal of a portion of an APASCI file (a) and corresponding telephone noisy version (b).



To take into account several noise levels, 5215 SNR values (the size of APASCI), ranging from 10 dB to 30 dB and distributed according to a Gaussian probability density function (pdf), with mean 20 dB and standard deviation 5 dB, were generated.

4. DIALOGUE ARCHITECTURE

Our dialogue architecture can be used for information access in restricted domains. As previously said, a train timetable inquiry prototype was developed and used to collect 70 human-machine interactions.

The main features of the dialogue architecture are the following.

- Speaker independent, continuous speech recognition and understanding: recursive transition networks [5] are used to represent, at the same time, both the language model, which limits the search space during recognition, and the basic concepts to understand.
- Easy portability to different restricted domains: all the data related to a given application are part of a description which is fed by a *dialogue engine*. The dialogue strategy is also included in the description and can be shared among different applications.
- Mixed initiative: since all the sub-grammars of the application are always active, the user can give new information or can switch to a new sub-domain in each moment of the interaction, even during confirmations.

The basic idea, underlying the dialogue architecture, relies upon the dynamic language models described in [5], with which basic concepts can be modelled by some sub-grammars (either regular or context-free) that can be combined, run-time, to form a bigger language model, to be used for recognising a complete sentence. Furthermore, words belonging to a sub-grammar can be labelled, easily extracted from the stream, and associated to the corresponding basic concept. In this way there is no need to parse the recognised sentence, because the relevant information is already labelled, so that some minor text processing is required (for example for dates and hours).

The architecture of the system includes a dialogue engine which has to interpret the description of an application, which is both declarative (for what concerns basic concepts) and procedural (for the definition of the actions that must be executed in some dialogue states). Each basic concept has associated a set of features, that defines how it will be used during the user interaction. Concepts are structured into contexts, associated to proper sub-grammars; this allows a bottom-up context switching based on the user's sentences.

The train timetable domain is a task having a vocabulary of about 900 words. The test set is formed by 70 dialogues; all of them have been annotated. Each dialogue consists, on average, of about 13 sentences, for a total of 925 sentences (1149 basic concepts). Sentences have been manually checked and transcribed, both in terms of pronounced words and in terms of basic concepts expressed. Performance on this data will be reported in terms of concept accuracy. A dialogue sample (in Italian) is reported, showing system prompts, user replies and recogniser outputs. Useful concepts are identified by their uppercase labels.

SYS: dica la stazione di partenza.
USER: eeeh dovrei andare a vercelli e volevo partire sabato 14 cioe' sabato prossimo circa verso le 9
REC: dovrei andare a (ARRCITY(vercelli)ARRCITY) volevo partire (STARTDAY(sabato 14)STARTDAY) C e (STARTDAY(sabato prossimo)STARTDAY) circa (STARTHOURS(verso le 9)STARTHOURS))DEBUG2)
SYS: vuole partire verso le ore 9 il giorno sabato 14 novembre e arrivare a vercelli ?
USER: si'
REC: (YESNO(si')YESNO)
SYS: dica la stazione di partenza.
USER: da trento
REC: ma (STARTCITY(trento)STARTCITY)
SYS: vuole partire da trento ?
USER: si' grazie
REC: (YESNO(si')YESNO) grazie
SYS: Sono stati selezionati 10 treni da trento a vercelli con partenza sabato 14 novembre verso le ore 9. quali treni vuole selezionare?
USER: eeeh mmmh
REC: ma 'sto
SYS: non ho capito. quali treni vuole selezionare?
USER: il numero 2
REC: i (CHOICE(1)EXACT) (EXACT(2)EXACT))CHOICE)
SYS: treno 1 : treno eurocity numero 546 con trasporto biciclette , partenza alle 7 e 17 , arrivo alle 12 e 52. treno 2 : treno eurocity numero 3274 con trasporto biciclette , partenza alle 7 e 40 , arrivo alle 12 e 47. posso fare altro per lei?

5. EXPERIMENTS AND RESULTS

In the experiments a set of 49 phone HMMs, derived from the SAMPA alphabet, has been used. A background noise HMM, plus three other HMMs related to: noises of various types, breaths or coughs and hesitations have also been introduced. In general, HMMs are defined by a three state left to right topology, without skips among states. A two state left to right topology has been adopted for bursts, liquids and semivowels, while background noise is modelled by a single state topology. Output pdf are defined by mixtures of 16 Gaussian functions, having diagonal covariance matrices.

Acoustic features are obtained by applying Linear Predictive Coding (LPC) analysis to 20 *ms* Hamming windows, at a rate of 10 *ms*. In each speech frame 12 cepstral coefficients, the log-energy and the corresponding first and second order time derivatives have been evaluated, thus obtaining an observation vector of dimension 39. To compensate for transmission channel effects Cepstral Mean Subtraction (CMS) has been used. More specifically, the mean vector of cepstral coefficients, evaluated over a sliding window of duration 1 *second*, has been subtracted to the coefficients themselves. The maximum of the log-energy, also evaluated over a sliding 1 *second* window, has been subtracted to the log-energy.

Several lists of files have been used for system training: for each of them corresponding performance has been evaluated. The training lists are shown in Table 1, where columns correspond to available data sets (**AF** means band filtered APASCI, **AS** means simulated APASCI, **P1** means PHONE1 and **P2** means PHONE2) and crosses mean that the data set has been included in the

training list of the corresponding row. As previously seen, HMMs are always bootstrapped from the band filtered version of APASCI (**AF**), hence the corresponding data set is always included in the training lists of Table 1.

Table1. Training lists used in the experiments.

	AF	**AS**	**P1**	**P2**
AF	X			
AFAS	X	X		
AFP1	X		X	
AFASP1	X	X	X	
AFP1P2	X		X	X
AFASP1P2	X	X	X	X

In Table 2. results are reported for both the test set **CARITRO** and **FERROVIE**. For **CARITRO**, performance are given in terms of Word Accuracy (WA). In this case, as previously seen, each utterance contains one of the names of the people or branches of the Bank, preceded or followed by typical expressions for asking for the name itself. The recognised name is extracted from the string provided by the recogniser and is compared with the correct name. In this way, no insertion or deletion errors are possible for this task. Note the effectiveness of using the simulated data when no telephone data are available for retraining (first and second rows of the Table). Effectiveness decreases as the quantity of telephone data used for retraining increases (last two rows of the Table).

Table 2. Performance, obtained on the CARITRO and FERROVIE tasks, using different training lists.

	CARITRO	**FERROVIE**
AF	73.30%	60.20%
AFAS	82.40%	61.00%
AFP1	87.30%	71.30%
AFASP1	89.20%	71.40%
AFP1P2	88.90%	73.40%
AFASP1P2	89.90%	73.60%

For the task **FERROVIE** performance is given in terms of semantic accuracy. A concept error can be: an insertion (a concept arises from the recognised sentence but was not expressed), a deletion (the system skips a valid concept), or a substitution. Since a concept can be defined as a pair <*slot*, value>, the latter error type include both cases in which the value is wrong (e.g. <*starting-city*, Roma> recognised instead of <*starting-city*, Verona>) and cases in which the value is correct but the slot is wrong (e.g. <*starting-city*, Roma> recognised instead of <*arrival-city*, Roma>).

The effectiveness of using simulated data is not tangible for the task **FERROVIE**. Probably the grammars used in the dialogue model mask, in some manner, the effects of the acoustic models. Furthermore, while some work initially was done for tuning some language model parameters (i.e. penalties have been introduced for rejecting some sub-networks), in order to maximise the performance of the baseline models (**AFP1P2**), the same was not done for the models derived from the other training lists.

6. CONCLUSION

In the paper we have described our approach for generating noisy telephone data, starting from a clean wide band speech database (APASCI). Also, a brief description of the activity we are conducting on spoken dialogue management has been given. The use of simulated data has been tested on two different tasks. One (**CARITRO**) is more *acoustic;* significant performance improvements have been measured on it, especially when few telephone data are available for training. The other (**FERROVIE**) is more l*inguistic*; in this case no significant improvements in recognition accuracy have been measured.

Future works will address some topics for improving both the acoustic and linguistic robustness of the actual dialogue architecture. In particular, flexible keyword verification methods, to avoid user's confirmations, as well as statistical language models (i.e. N-grams), specific of each basic concept, are going to be developed.

7. REFERENCES

[1] Angelini B., Brugnara F., Falavigna D., Giuliani D., Gretter R., Omologo M. (1994), Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. *Proceedings of ICSLP*, Vol. 3, pp. 1391-1394, Yokohama, 1994.

[2] Falavigna D., Gretter R. (1998), Telephone Speech Recognition Applications at IRST. *Proceedings of IV IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 27-30, Turin, 1998.

[3] Suzuki Y., Asano F., Kim H., Sone T. (1995), An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. In: *Journal of the Acoustic Society of America*, Vol. 97, No. 2, February 1995, pp. 1119-1123.

[4] Giuliani D., Matassoni M., Omologo M., Svaizer P. (1999), Training of HMM with Filtered Speech Material for Hands-Free Recognition. In: *Proceedings of ICASSP*, Vol. 1, pp. 449-452, Phoenix, 1999.

[5] Brugnara F., Federico M. (1997), Dynamic Language Models for Interactive Speech Applications, *Proceedings of EUROSPEECH*, Rhodes, 1997.