

From generic to task-oriented speech recognition : French experience in the NESPOLE! European project

Dominique Vaufreydaz, Laurent Besacier, Carole Bergamini, Richard Lamy

▶ **To cite this version:**

Dominique Vaufreydaz, Laurent Besacier, Carole Bergamini, Richard Lamy. From generic to task-oriented speech recognition : French experience in the NESPOLE! European project. ITR-Workshop on Adaptation Methods For Speech Recognition, Aug 2001, Nice - Sophia Antipolis, France. pp. 179-182. inria-00326171

HAL Id: inria-00326171

https://inria.hal.science/inria-00326171

Submitted on 2 Oct 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

From generic to task-oriented speech recognition : French experience in the NESPOLE! European project

D. Vaufreydaz, L. Besacier, C. Bergamini, R. Lamy

CLIPS-IMAG Laboratory, Joseph Fourier University
B.P. 53, 38041 Grenoble cedex 9, France
(dominique.vaufreydaz, laurent.besacier, carole.bergamini, richard.lamy)@imag.fr

Abstract

This paper presents CLIPS laboratory activities in speech recognition related to language model adaptation and acoustic model adaptation in the NESPOLE! European project. ASR system needed to be adapted in two ways. The language model had to deal with task specific vocabulary and the acoustic model had to be robust to VoIP (Voice over IP) speech.

It was shown that Internet, as a very large source of text, can be a very interesting database for spoken language modelling adaptation. The influence of different VoIP codecs on the performance of our speech recognition engine was investigated and a new strategy was proposed to cope with degradation due to low bitrate coding. The acoustic models of the speech recognition system were trained with transcoded speech. Results have shown that this strategy allows to recover acceptable performance for the NESPOLE! project context.

1. Introduction

The NESPOLE![1] Project is a common EU NSF funded project exploring future applications of automatic speech to speech translation in e-commerce and e-service sectors [1]. Spoken languages involved in this project are Italian, German, English and French. Partners of the project are *ITC/IRST* from Trento (Italy), *ISL Labs*. from *UKA* (Karlsruhe, Germany) and *CMU* (Pittsburgh, USA), *Aethra* (an italian company specialised in videoconferencing software), *APT* a tourism agency in the Trentino area (Italy) and finally *CLIPS* laboratory (Grenoble, France).

The scenario for the first showcase of NESPOLE! involves an Italian speaking agent, located in a tourism agency in Italy and a client located anywhere (English, German or French speaking) using a simple terminal (PC, sound and video cards, H323 videoconferencing software like NetMeeting™). This choice is related to present available technology, in the near future the third generation cellular can be also used as terminal.

The client wants to organise a trip in the Trentino area, and refers to APT (the tourism agency) web pages in order to get information. If the client wants to know more about a particular topic or prefers to have a more direct contact, a speech to speech translation service allows him to interact in his own language with an APT Italian agent. A videoconferencing session can then be opened between client and agent and the dialog starts between them.

In this project, the scientific and technological research issues intended to be addressed in order to improve current experimental speech-to-speech translation (STST) systems, are: robustness, scalability, cross-domain portability and multimodal interaction with multimedia content.

This paper particularly deals with adaptation of the automatic speech recognition module which is an essential part of the complete speech to speech translation chain. ASR system needs to be adapted in two ways. Indeed, the language model must deal with task specific vocabulary ; this will be presented in *section 2*. Moreover, the acoustic model needs to be adapted to be robust to VoIP (Voice over IP) speech ; this will be presented in *section 3*. Experimental results will be described in *section 4* and finally *section 5* will draw conclusions and give some perspectives.

2. Language model adaptation

2.1. Vocabulary definition

2.1.1. Specific task vocabulary

The first step to achieve a language model is the vocabulary definition. We must define which words are used in the context of the application. To do that, a large data collection was performed in summer 2000 in each NESPOLE partners language [2].

191 dialogs in 4 languages were recorded, involving a tourist client that request information to an agent located in APT (tourism agency) in Italy. 5 different scenarios were developed (winter accommodation, all-included tourist package, summer vacation in a park, castle and lake tours, and looking for folklore and brochures).

31 French dialogs were recorded using the videoconferencing software NetMeeting. Manual transcriptions of the 31 recorded French dialogs were used to determine a task vocabulary (related to the scenarios defined) of 2056 words.

2.1.2. CStar vocabulary

To complete this vocabulary, we used the one defined for the CStar-II[2] project which also concerns tourism information but in a more constraint form (not all the topics used in NESPOLE!, less activities, etc.). It was defined by transcribing Wizard of Oz experiments conducted at our laboratory. A couple of words have been added empirically during the preparation of the CStar-II demonstration. The new vocabulary, obtained by the union of the two specific lexicons, grows up to 2500 words.

[1] http://nespole.itc.it/

[2] http://www.c-star.org/

2.2. Language model adaptation

2.2.1. Vocabulary enhancement

As we know, spoken language is very difficult to model because of specific phenomena like hesitations, speech repairs, etc. Moreover, we can not prevent users from using out of vocabulary (OOV) words. Thus, we should enlarge the based vocabulary in order to deal with these facts.

Adding frequent French words

We decided to add the most frequent French words to complete the word list. We computed a word count on our last Web Collection *WebFr4*. It is a very large corpus containing a few less than 6 millions Web Pages representing 44 Gigabytes. These documents are gathered by our bot Clips-Index[1] on French speaking domains except Switzerland (not enough percentage of French documents) and Canada (too far from us on the network).

In order to check for French words in *WebFr4*, we need an exhaustive list of them. The maximal list of French lexical words is constructed with two French lexicons, BDLex [3] and the ABU[2] dictionaries, and consists of more than 400,000 forms. With this list, we can count in the text extracted from *WebFr4*, more than 3 billions words. We obtained the French coverage of this corpus. We found a few less than 200,000 different lexical forms. We sorted them according to their frequency and complete the previous vocabulary to obtain 20,000 entries for the speech recogniser.

Adding compounds words

The motivation for making compound words is that there is a word insertion penalty in the general language modelling equation in order to regulate the number of words in the recognition hypothesis. So, “il y a” (French equivalent of “there is” in English) is considered by the model as 3 words, but these 3 words represents only 4 phonemes and a longer word, which is somehow phonetically close, can be, under some assumptions, wrongly chosen.

The first idea can be to set an inverse value for the penalty that becomes thus a bonus. But in this case, the system gives priority to short words. In the extreme case, if this bonus is too high, the system only produces an output hypothesis containing uni-phonemes entries, like enumeration letters for example. The second way to deal with short words is to compute compound words that are treated in the language model as one word. There are two benefits in this method. Firstly, there is no biased usage of the word penalty of the recogniser. Secondly, in this case, we increase the context taken into account in the language model. For a *n*-gram model, when one or several of the items is a compound word, the learned context increases from *n* to the real length of the sequence represented in the *n*-gram. For example, the 3-gram “il_y_a de_la vie” can be considered as a 6-gram.

In the state of the art, there are several methods to compute compound words. Most of them are data-driven and work on the acoustic forms of words and some spoken phenomena like plosive deletion or palatization [4]. As we want to use statistical language modelling and because we get a very large training corpus, we decided to use only statistical count of compound words in the text extracted from *WebFr4*. We limited this count on combinations of words that not exceed 5 letters and 3 phonemes in the actual vocabulary. Then, we computed 5-grams on our Web corpus and kept only the best combinations of each level (i.e. the best bigrams, the best trigrams and so on). Using thresholds, we kept at least 524 compound words added to the lexicons of the recogniser. We can note that, with the same thresholds and with the 400,000 words lexicon, we found more than 1,700 French compound words. We can see similar approach in [5].

The final vocabulary for our recogniser contains 20,524 entries and the phonetic dictionary, containing these words and their phonetics variants, contains 27,117 entries.

2.2.2. Computing Language Model

The last step is to calculate the language model. We use 3-grams models computed on our Web corpus using the “minimal block length” method described in [6]. With this lexicons and a minimal block length of 5, we filtered *WebFr4* and obtained to learn our language model a 1,587,142,200 words corpus. The Resulting language model contains 1,960,813 bigrams and 6,413,376 trigrams.

The recognition results given in *section 4* are obtained using this language model.

3. Acoustic model adaptation

Different strategies can be proposed to access a distant speech recognition server:

-Recently, some solutions were proposed where speech analysis (the extraction of acoustic features for speech recognition, which is a low-cost task) is performed on the client terminal and the acoustic parameters are directly transmitted to the distant speech recognition server. The expression *distributed speech recognition*[4] is generally used to define this type of architecture [7,8]. In this kind of architecture, the terminal needs to be able to extract the adequate acoustic parameters needed at the input of the speech recognition server.

-Sometimes, one may want to access to a speech recognition service from a very basic terminal which is just able to capture and transmit speech. In this case, the speech signal is transmitted from the client terminal to a distant speech recognition server. Coding of the speech signal is generally necessary to reduce transmission delays (GSM, MPEG or G7XX coding, depending on the client terminal used). This is the case in NESPOLE! scenario where a client can use the speech translation service from anywhere. Thus, speech needs to be achieved from the client terminal to the distant speech to speech translation modules, the first module being speech recognition. As a H323 videoconferencing software is used, the speech signal is coded to reduce transmission delays (G711, G723.1, G728 codecs…). As transcoding (the process of coding and decoding) modifies the speech signal, it is likely to have an influence on speech recognition performance if acoustic models are not adapted. We first tried to verify this hypothesis by performing speech recognition experiments on transcoded speech. Then we proposed a solution to cope with "transcoding" effects.

---

[1] http://clips-index.imag.fr/

[2] http://abu.cnam.fr/

[3] http://www.c-star.org/

[4] http://www.icp.inpg.fr/ELRA/aurora.html

The solution proposed is to train the acoustic models with transcoded speech. For this, the Bref80 [9] database (that we use for training our acoustic models) was transcoded via different coders in order to train one acoustic model for each speech coder used in our application. To know which speech coder is used on the user terminal (and thus to know which acoustic model must be used) the information concerning the coder used by the client is encoded in the data stream transmitted from the client to the speech recognition server.

4. Experiments

4.1. Databases

4.1.1. Training databases

We used a corpus which contains 12 hours of continuous speech of 72 speakers extracted from Bref80 [9]. Two transcoded (coded then decoded) versions of this Bref80 corpus (originally 256 kbits/s) were also obtained. The audio codecs used in H323 standard are G711, G722, G723, G728 and G729. We used in our experiments the codec which has the lowest bit rate : G723.1 (5.3 kbits/s), and the one with the highest bitrate : G711 (64 kbits/s : 8 kHz, 8 bits). Source code of these coders can be found on ITU-T web site[1] but in fact for G711, you just have to downsample to 8khz and requantize to 8 bits in order to simulate a G711 degradation. In the results we will refer to the training databases *Bref80*, *Bref80-G711* and *Bref80-G723.*

4.1.2. Test databases

Cstar

The first test database is made up of 120 recorded French sentences focused on reservation and tourist information task. These sentences (originally 256 kbits/s bitrate) were also transcoded via a G711 (64 kbits/s) coder and G723.1 (5.3 kbits/s) coder. We will refer to the test databases *Cstar*, *Cstar-G711* and *Cstar-G723* in the experiments. This database is rather easy for our task since the sentences were pronounced carefully by speakers used to speaking to a speech recognition system.

Nespole

This second test database is made up of 77 sentences (extracted from transcriptions of NESPOLE dialog database [3]) recorded on a client terminal and transmitted through the network with NESPOLE hardware architecture. On the distant site connected to the client terminal, VoIP G711 speech was collected. Thus we had for testing 77 speech signals transmitted through the network with G711 coding (will be referred to *Nespole-G711*). This small database represents the speech quality that can be encountered in NESPOLE environment; however for *Nespole-G711*, only the effect of speech coding is measured since the speech was transmitted through a LAN (Local Area Network) with almost no data packet loss. Experiments on data transmitted through a WAN (Wide Area Network) with packet loss are currently performed but are not provided in this paper.

[1] http://www.itu.int/publications/itu-t/itutrec.htm

4.2. Speech Recognition System

Our continuous French speech recognition system RAPHAEL uses Janus-III toolkit [10] from CMU. The context dependent acoustic model (750 CD codebooks, 16 gaussians each) is learned on Bref80 and its "transcoded" versions. The system uses 24-dimensional LDA features obtained from 43-dimensional acoustic vectors (13 MFCC, 13 ΔMFCC, 13 ΔΔMFCC, E, ΔE, ΔΔE, zero-crossing parameter). The vocabulary and the trigram language model used are those obtained as described in *section 2*.

4.3. Results

We first calculate word recognition performance (WR% is equal to 100 – WER%) for the *Cstar* test databases using different acoustic models learned on different training databases. In order to be compared to an acoustic model learned on the original version of *Bref80*, signals from *Cstar-G711* and from *Cstar-G723* were upsampled to 16khz before the test (first line of the table).

test signals	*Cstar*	*Cstar-G711*	*Cstar-G723*
training signals	WR%	WR%	WR%
Bref80	92.28%	72.01%	63.64%
Bref80-G711		91.89%	91.2%
Bref80-G723		91.11%	91.2%

Table 1 : speech recognition performance (word recognition) versus data quality on *cstar* DB

First line of Table 1 shows that a degradation of recognition performance is observed due to speech compression. Training an acoustic model on a database coded with a particular coder allows to recover this performance loss. However, since we do not see much difference between G711 and G723 performance (second and third line of Table 1), it seems that degradation observed on the first line is mostly due to frequency bandwidth reduction that occur when passing from original 16khz signals to G711 or G723 transcoded signals. So G723 which is a very low bitrate coder (5.3 kbits/s) has an influence on performance comparable with G711 coder.

In the second experiment, we calculated the performance obtained with our speech recognition on *Nespole-G711 DB.* The result presented in Table 2 shows that our adapted acoustic and language models allow to obtain acceptable performance in the context of the NESPOLE project.

test signals	*Nespole-G711*
training signals	WR%
Bref80-G711	82%

Table 2 : speech recognition performance (word recognition)

5. Conclusions and Perspectives

We have presented the CLIPS activities in speech recognition related to language model adaptation and acoustic model adaptation in the NESPOLE! European project.

It was shown that Internet, as a very large source of text, can be a very interesting database for spoken language

modelling that continues to grow every day. Moreover, using web pages, we can cope with several problems in language modelling for speech recognition like compounds words and vocabulary modification. In the future, we intend to work on multilingual language modelling using the same kind of techniques used for French, adding multilingual text alignment from Internet.

We have also investigated the influence of different VoIP codecs on the performance of our speech recognition engine. A new strategy was proposed to cope with degradation due to low bitrate coding. The acoustic models of the speech recognition system were trained with transcoded speech. Results have shown that this strategy allows to recover acceptable performance. However, we did not observe a very significant difference in performance between G711 and G723.1 coders although G723.1 has a much lower bitrate than G711.

As a perspective, a filtering strategy could also be proposed to cope with transcoding degradation : instead of degrading the acoustic models to reduce mismatch between training and test data, we could keep the acoustic model trained on clean speech and learn a linear transformation f between "clean" and "degraded" signals. During the recognition stage, the inverse transformation f^{-1} would be applied to the degraded test signal to reduce mismatch between training and test.

Another perspective, not far from the work presented in [7] is the idea of extracting acoustic vectors for recognition locally on the client terminal, and to classify the feature space thanks to Vector Quantization (VQ) methods. Thus, instead of transmitting coded speech or acoustic vectors, we could transmit the only class number of each vector. The data transmission rate could be thus decreased. For instance, if we quantize the acoustic features with a 2^{16} codebook size, and if 100 feature vectors are extracted every second (10ms frame rate) then one just have to transmit 100 values coded on 16 bits per second. The bitrate is then decreased from 64 kbits/s (for G711) to 1.6 kbits/s. At the reception, we replace the class number with the centroid representing this class.

To evaluate the feasibility of this, and the sufficient number of codebook needed, a first experiment has been done. The 43-dimensional feature vectors extracted from the 120 signals of *Cstar* DB were quantized with different codebook size and transmitted to the recognition system. The chosen VQ algorithm was the binary split algorithm [11]. The graph on Figure 1 shows the word recognition rate versus the codebook size. The reference word recognition rate when the acoustic vectors are not quantized is represented with the horizontal line.

The results show that acceptable performance can be obtained with more than 512 codebook size. This kind of architecture is thus very promising since it allows to keep very good speech quality without speech coding degradation, while keeping a low bitrate for transmission. The only problem is that the terminal on which features are extracted needs to be able to extract the adequate acoustic parameters needed at the input of the speech recognition server. Quite surprisingly, the performance is slightly higher than the reference for more than 1024 codebook size. It seems that in this case, VQ acts as a kind of filtering process which suppress some disturbing information.

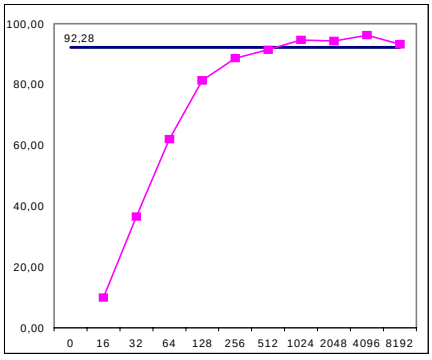


Figure 1 : speech recognition performance (word recognition) versus codebook size on *cstar* DB

6. References

[1] Lazzari G., Spoken translation : challenges and opportunities, *ICSLP'2000*, Beijing, China. Oct. 16-20, 2000, vol 4/4 : pp. 430-435

[2] Burger, S., Besacier, L. Metze, F., Morel, C., Coletti, P., The NESPOLE! VoIP dialog database, *Eurospeech 2001*. Accepted.

[3] Pérennou G., De Calmès M., "BDLEX lexical data and knowledge base of spoken and written French", *European conference on Speech Technology*, pp 393-396, Edinburgh (Scotland), September 1987.

[4] Saon G., Padmanablan M., *Data-driven Approach to Designing Compound Words for Continuous Speech Recognition*, ASRU'99, pp. 261-264, Keystone, Colorado (USA), December 1999.

[5] Beaujard C., Jardino M., *Language Modelling Based on Automatic Word Concatenations*, Eurospeech'99, pp. 1563-1566, Volume 4, Budapest (Hungary), September 1999.

[6] Vaufreydaz D., Akbar M., Rouillard J., *Internet Documents: A Rich Source for Spoken Language Modelling*, ASRU'99, pp. 277-280, Keystone, Colorado (USA), December 1999.

[7] Pearce, D., Motorola Labs, "An Overview of ETSI Standards Activities for Distributed Speech Recognition Front-Ends", *AVIOS 2000 : The Speech Applications Conference*, May 22-24, 2000 San Jose, CA, USA.

[8] Zhang, W., He, L., Chow, Y., Yang, R., Su, Y., "The Study on Distributed Speech Recognition System", *ICASSP 2000 International Conference en Acoustic Speech & Signal Processing.* June 5-9 2000, Istambul, Turkey.

[9] Lamel, L.F., Gauvain, J.L., Eskénazi, M. "BREF, a Large Vocabulary Spoken Corpus for French", *Eurospeech*, Gênes, Italy, Vol 2, pp. 505-508, 24-26 September 1991.

[10] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. "Recent Advances in JANUS : A Speech Translation System". *Eurospeech*, 1993, volume 2, pages 1295-1298.

[11] Rabiner L., Juang B-H., "fundamentals of speech recognition " ed. Prentice Hall PTR, 1993, 507 p.