Using prompts to produce quality corpus for training automatic speech recognition systems

Benjamin Lecouteux, Georges Linarès

Using prompts to produce quality corpus for training automatic speech recognition systems

Benjamin Lecouteux and Georges Linarès

***Abstract*— In this paper we present an integrated unsupervised method to produce a quality corpus for training automatic speech recognition system (ASR) using prompts or closed captions. Closed captions and prompts do not always have timestamps and do not necessarily correspond to the exact speech. We propose a method allowing to extract quality corpus from imperfect transcript. The proposed approach works in two steps. During the search, the ASR system finds matching segments in a large prompt database. Matching segments are then used inside a Driven Decoding Algorithm (DDA) to produce a high quality corpus. Results show a F-measure of 96% in term of spotting while the DDA corrects the output according to the prompts: a high quality corpus is easily extracted.** [1]

***Index Terms*— speech recognition, closed captioning, corpus building, automatic segmentation**

I. Introduction

The training of an automatic speech recognition system (ASR) requires large amounts of exact annotated speech. The transcription task is expensive and takes a lot of time. In some situations imperfect transcripts like journalist prompts, closed captions or abstracts are available. This material is available in large quantities.

However, these transcripts present two issues: the distance compared to the audio stream and the lack of timestamps. Various approaches propose to use imperfect transcripts for unsupervised ASR training (section II-D). But existing methods are not integrated and have shortcomings: processes are iterative and take a lot of computing time; the lack of timestamp is forgotten. Moreover existing methods do not use all the potential of imperfect transcripts.

The first part of this paper is dedicated to the related work on these issues: the prompts quality, methods to perform ASR alignment with imperfect transcripts, the automatic imperfect transcript segmentation, and finally how to use them for training an ASR system.

In a second part, we describe an integrated approach allowing us to solve the two main approximated transcription issues:

- In section III-A we describe the driven decoding algorithm which allows us to drive an ASR according to a transcript. DDA allows to correct on the fly an approximated transcript.Then we present some DDA experiments, results and the DDA ability to exploit dynamically imperfect transcripts.
- The section III-B presents the "spotting text island algorithm" to select segments in real-time into a large prompt database. This algorithm allows to synchronize prompts with the ASR system.The experiments are carried out on the RTBF and ESTER databases.

The section III-C presents the text spotting integration into the driven decoding algorithm. We discuss how this method makes it possible to build high quality corpora. Finally the last section presents the conclusions and future works.

[1]This research is supported by the ANR (Agence Nationale de la Recherche), AVISON project.

B. Lecouteux and G. Linarès are with the Laboratoire Informatique d'Avignon (LIA), University of Avignon FRANCE (e-mail : benjamin.lecouteux, georges.linares@univ-avignon.fr)

II. Related work

A. Quality of prompts or closed captions

P. Cardinal presents in [1] the most common problems with journalist prompts: journalist stories associated to speech are imperfect. Prompts are not always respected, sentences are forgotten or inserted, and other sentences are pronounced with some word variations because texts are often just a guideline for the journalist. In the case of closed captions the WER are 10% to 20% compared to the exact transcript [2]. Experiments computed by M.J. Witbrock [3] have shown that approximately 16% of the words are incorrectly transcribed compared to the exact transcript. Moreover he shows that using transcripts *a-posteriori* is not always relevant: experiments with a trigram language model derived from a correct transcript showed a high word error rate.

B. Automatic closed caption alignment

Different methods are proposed to align closed-captions with the output of an ASR. In [4], P.J. Moreno & al. propose to align long audio documents with their exact transcripts within the framework of automatic indexing of multimedia documents. His method is based on the search of anchor points which are isolated from the extracted segments with a high correlation between the *a-priori* transcript and the automatic transcript. However this method is only applicable to low-error texts.

In [1] P. Cardinal & al. aligns ASR outputs with imperfect texts by computing their edit-distance which is the minimal cost to obtain a final sentence. To compute minimal cost, M. Mohri uses FSTs [5]. The two strings are represented by transducers $T_{outputASR}$ and $T_{closedcaption}$. An edition funtion T_{edit} is associated to each closed caption. Then the set of all alignments is computed by :

$$T_{aligns} = T_{outputASR} \oplus T_{edit} \oplus T_{closedcaption} \qquad (1)$$

The best path is computed by performing a best path search :

$$BestPath = BPS(T_{aligns}) \quad (2)$$

This operation is performed for each potential closed-caption. A module rejects bad alignments. Good alignments are used for training acoustics models. However this is an *a-posteriori* method: more than one pass must be performed and the alignment is heavy in terms of calculation time. Moreover bad decoded segments can not be aligned.

In [3], M.J. Witbrock uses timestamps and aligns matching closed-captions with classic dynamic time warping (DTW, [6]) algorithm. However this is a batch method with the same limitations that method proposed in [1].

P. Placeway proposes a more integrated method in [2]. He proposes a translation model for mapping caption sequences to word sequences which updates the language model. The translation model (figure 1) is a Markov chain where arcs represent deletions, insertions or substitions: this model corresponds to the string edit distance with a DTW. The model is integrated in the decoder but timestamps are essential to select the good caption model. These experiments are carried out using a synchronous decoder (Sphinx-3, [7]).

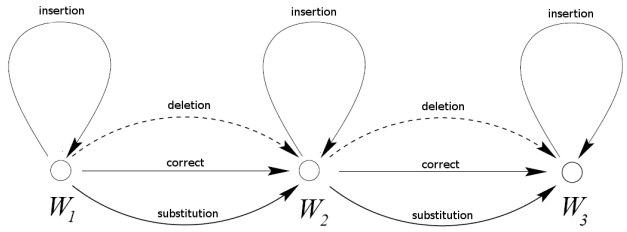

Fig. 1. The translation model (a Markov chain) where arcs represent deletions, insertions or substitions

Previous methods are limited to low WER transcripts. In a recent paper A. Haubold proposes an approach for the alignment of speech to highly imperfect transcripts in [8]. His approach is similar to the previous, but temporally unaligned text is converted to phonemes. Then the alignment is performed by using an edit-distance similar to the alignment between two DNA sequences. This obtained promising results.

C. Automatic segmentation on partial transcriptions

The goal of segmentation is to select the segments in a (large) prompt database which match the current speech. In [9] the authors correct the approximate time marks from the closed-caption transcriptions. They align the closed-caption with the automatic audio transcription. Then the algorithm assigns timestamps to closed captions according to the automatic transcription. However, in this approach, time marks are available in closed captions. The work presented in [10] also uses available closed-captions with their boundaries.

In [1], P.J. Moreno proposes a method allowing the selection of segments in a large database. A disadvantage of this approach is the computing time: all potentials segments must be tested. This approach presents a segmentation accuracy of 80% with a false reject rate of 30%.

Some algorithms adapted to find a local segmentation have been developed: the Smith-Waterman [11] algorithm which is designed to perform local sequence alignment; originally for determining similar regions between two nucleotide or protein sequences. The Smith-Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure to find sub-sequences. However the algorithm requires a lot of time and memory.

The problem can be tackled as textual information retrieval. In text retrieval, the problem is to find documents meeting the user's information need. The vector model is the most used model for information retrieval. With this model documents are represented in a space D whose dimensions are the words composing the documents. Words are extracted from the documents after stripping stop words and stemming them [12]. Queries are represented in the same space as documents, like the documents. The tf x idf (term frequency x inverse document frequency) is often used to get an estimation of the information carried by a word [13]. The similarity between queries and documents are computed (The cosine similarity is widely used : it is the cosine of the angle between the document vector and the query vector) and documents ranked by the similarity measure. This operation is fast (section III-B) and allows us to retrieve a matching segment from a large document.

D. Using prompts for training an ASR

One of the main interest in improving corpus quality is the training of acoustic models for ASR systems.

In [10], L. Lamel proposes a recursive method for acoustic model training using low quality transcribed databases. This approach consists of three steps :

- Decode the training database automatically.
- Find matched segments between approximated transcripts and automatic decoding.
- Matching segments are used for acoustic model re-estimation.

This method was tacked by M.J. Witbrock [3]. He uses teletext/closed-caption for re-training acoustic models based on television input. The ASR is used to find matching closed-captions. These segments are then used to retrain acoustic models. In [14] the authors investigate the use of closed-captions for MMI (Maximum Mutual Information) or MPE (Minimum Phone Error) discriminative training with a similar scheme.

Another approach proposed by P. Placeway [2] is to estimate a language model with prompts or closed-captions. The estimated model is interpolated with a generic language model in the ASR. This technique improves the results, but some subtitles information is drowned in the data quantity. Moreover, P. Placeway included a closed caption model into the beam search: words matching the imperfect transcript are favored. This method is limited to closed-captions with timestamps.

In [15], the authors propose a light supervision method to acquire acoustic training data from speech having corresponding prompts. They estimate a biased language model using imperfect transcripts. The ASR output is aligned to the approximated transcripts and only matching words are selected for acoustic training. They obtain 13% relative error rate reduction with 702 hours added to the baseline (141 hours of training data). However the partial imperfect transcript information is not directly included in the decoder.

III. Approach

Our objective is to exploit imperfect transcripts like prompts or closed-captions when no timestamp information is available. The proposed method is integrated: matching segments are selected on the fly in a large database by the ASR system. Once matching segments are selected, they drive the ASR system.

Firstly we present the driven decoding algorithm (DDA) and secondly the fast-match to transcript island.

A. The Driven Decoding Algorithm

Previously, we proposed a Driven Decoding Algorithm (DDA) which is able to simultaneously align and correct the imperfect transcripts [16]. DDA works with SPEERAL [17], an asynchronous decoder derived from the A^* algorithm. The ASR generates assumptions as it progresses the phoneme lattice. For each new step, the current assumption is aligned with the approximated transcript (figure 2). Then a matching score α is computed and integrated with the language model:

$$\tilde{P}(w_i|w_{i-1}, w_{i-2}) = P^{1-\alpha}(w_i|w_{i-1}, w_{i-2}) \qquad (3)$$

where $\tilde{P}(w_i|w_{i-1}, w_{i-2})$ is the updated trigram probability of the word w_i knowing the history w_{i-2}, w_{i-3}, and $P(w_i|w_{i-1}, w_{i-2})$ is the initial probability of the trigram.

When the trigram is aligned α is at a maximum and decreases according to the misalignments of the history (values of α are determined empirically using a development corpus).

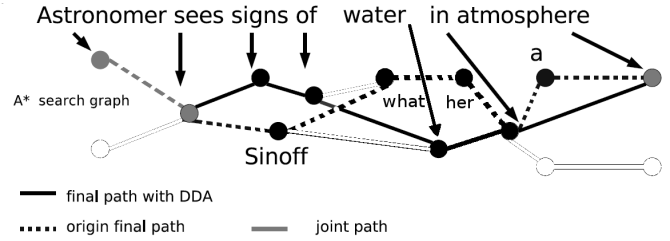

Fig. 2. The DDA mechanism drives the search by dynamically rescoring search function according to the alignment scores.

1) Experiments: In the first experiment, approximated transcripts are already segmented. The alignment is integrated into the ASR, and based on the Swith-Waterman algorithm. DDA experiments were carried out with the "Broadcast news" system developed by the LIA for the French evaluation campaign ESTER [18]. The system is assessed on 3 hours of radio broadcast extracted from the ESTER development corpus (France Inter 1, France Inter 2, France Info). Imperfect transcripts are made by adding errors manually in the exact transcript: 10% WER in France Inter shows, and 20% WER in France Info show.

2) Alignment with exact transcripts: Preliminary experiments with exact transcripts allow us to evaluate the DDA potential (table I). The baseline system is 22.7% WER. Using a language model estimated on the transcript reduces the WER to 5.2%. However better results are found when using alignment during the graph search (section III-A): 3.7% of WER. Our experiments show that using the estimated language model with the dynamic alignment does not improve the system. The minimal WER for a method re-estimating the concurrent hypothesis without modifying the content of the hypothesis stack is 3.7%.

	WER
FrInter 1: ML-G alone	22.7%
FrInter 1: ML-TrEx alone	5.2%
FrInter 1: ML-G + ML-TrEx	10.8%
FrInter 1: ML-G alone + alTrEx	3.7%
FrInter 1: ML-TrEx alone + alTrEx	3.7%
FrInter 1: ML-G + ML-TrEx + alTrEx	3.7%

TABLE I

Experiments with exact transcript : ML-G is the generic language model, ML-TeEx is the language model estimate on the perfect transcript and alTrEx is an dynamic alignment to the perfect transcript during the graph search

3) Alignment with imperfect transcripts: Experiments with approximated transcripts show the same behavior (table II). Interpolated language models improve the baseline, but temporal information is lost. Using dynamic alignment, imperfect transcripts drive the decoder, and reciprocally, the decoder corrects the imperfect transcripts.

	WER
FrInter 1: ML-G alone	22.7%
FrInter 1: ML-TrErr alone	16.3%
FrInter 1: ML-G + ML-TrErr	15.2%
FrInter 1: ML-TrErr + alTrErr	9.9%
FrInter 1: ML-G + alTrErr	7.7%
FrInter 1: ML-G + ML-TrEr + alTrErr	7.2%

TABLE II

Interpolation of the generic language model (ML-G) with the model trained on the imperfect transcript (ML-TrErr - 10% WER)

4) Final results with DDA: In order to validate these results, we tested the system on a larger corpus. Two hours are processed using the same evaluation protocol described in the last sections. We observe that the gain in performance seems to be relatively independent from the quality of the initial transcript (table III).

These experiments show that the DDA is able to correctly align imperfect transcripts during the search. Moreover DDA dramatically improves the initial transcript: any imperfect information is exploited and corrected. Contrary to *a-posteriori* alignments DDA directly improves the decoder quality.

Shows	Baseline	Transcript	TDS
France Inter 1	22.7%	10.1%	7.2%
France Inter 2	21.1%	10.2%	7.7%
France Info	24.3%	20.3%	12.1%

TABLE III

WER OBTAINED BY THE BASELINE SYSTEM (*Baseline*), WER IN THE ORIGINAL TRANSCRIPT (*Transcript*), WER OBTAINED BY TRANSCRIPT DRIVEN SYSTEM (*TDS*)

B. Spotting transcript-Islands

Previously, the DDA algorithm was not able to align large segments in reasonable time. The Swith-Waterman algorithm complexity is $O(n, m)$ with n the number of query terms and m the number of document terms. In order to apply DDA to transcript-island spotting we add a fast-match process which aims to find on-the-fly the transcripts segments which are relevant to the current state of the search algorithm [19].

1) Fast-match transcript-island: The principle of the text island spotting (figure 3) is close to approaches used in the field of information retrieval. In our case, the hypothesis is a query searching a transcript-island. Search engines try to find the most relevant documents by comparing the query to the indexed collection of stored documents. As the current hypothesis is developped, a set of word clusters are built or updated. Clusters are the intersection between query and the large database. An adapted similarity measure is performed on each cluster and the most relevant are selected for alignment (based on the score being greater than a fixed *a-priori* threshold).

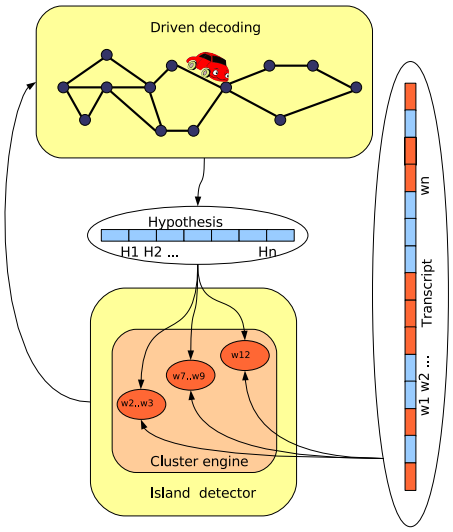

Fig. 3. Scheme of principle of spotter integration to Driven decoding. Transcripts-islands are detected by the spotter which computes an assumption-to-island matching score. According to it, the spotter sets the decoder in driven decoding mode.

2) Spotting exact transcript island on ESTER database: In the first experiment, we have tested the spotting algorithm with exact transcript. Moreover we have removed 50% of the transcription to test the spotting quality. Experiments are assessed on three hours extracted from the ESTER development corpus (France INTER, France INFO, RFI). During the search each assumption is presented as a query: the spotting algorithm proposes a corresponding segment in the database. The table IV presents the results. With perfect transcripts, the F-measure is 95.5%.

Radio station	Precision	Recall	F-measure	Seg. number
FrInter	90.7%	96.9%	93.7%	478
FrInfo	93.4%	89.7%	91.5%	468
RFI	98.8%	97.8%	98.4%	812
Mean	95.3%	97.3%	95.5%	1758

TABLE IV

SPOTTING ON THE ESTER DATABASE. EXPERIMENTS ARE PERFORMED ON 3 HOURS OF THE DEVELOPPEMENT SET, BY USING IMPERFECT TRANSCRIPTS OF ABOUT 10% WORD ERROR RATE. AS IN TABLE 1, 50% OF TRANSCRIPT SEGMENTS HAVE BEEN REMOVED FOR SPOTTING EVALUATION.

3) Spotting imperfect transcripts on ESTER database: In the second experiment, we have added errors into the transcripts (10%) to test the robustness of the presented spotting algorithm. Results are similar to the previous experiment (table V). In the table results on real condition with RTBF (Radio Télévision Belge Francophone, a Belgium broadcast news radio) database are added. This implies the algorithm is able to quickly find segments in spite of errors. It can be used to find segments in imperfect transcripts like prompts or closed captions.

Radio station	Precision	Recall	F-measure	Seg. number
FrInter	90.7%	96.9%	93.7%	478
FrInfo	93.4%	89.7%	91.5%	468
RFI	98.8%	97.8%	98.4%	812
RTBF	99.3 %	97.1 %	98.4 %	501
Mean	96%	95.8%	95.9%	2259

TABLE V

SPOTTING ON THE ESTER DATABASE. EXPERIMENTS ARE PERFORMED ON 3 HOURS OF THE DEVELOPPEMENT SET, BY USING IMPERFECT TRANSCRIPTS OF ABOUT 10% WORD ERROR RATE. THE RTBF SHOW IS AN EXPERIMENT IN REAL CONDITION

C. Transcript island Driven decoding

Finally we have tested the integration of spotting algorithm with DDA. Experiments are performed on 3 hours of the developement set, by using imperfect transcripts of about 10% Word Error Rate and 50% of text-segments have been removed for spotting evaluation. The table VI show the DDA search algorithm driven by the perfect transcript and DDA driven by imperfect transcripts.

This set of experiments showed that the spotting algorithm combined with DDA is able to produce a better transcript. Moreover we have the possibility to extract only aligned words. Aligned words are associated with a high level confidence score. Our approach allows us to take advantage

System	Baseline	DDA+IT	DDA+PT
INTER	22.6 %	17.9%	17.1%
INFO	23.4 %	21.7%	18.3%
RFI	27.2 %	23.0%	20.3 %
Mean	24.4 %	20.9 %	18.6 %

TABLE VI

WORD ERROR RATES WITH THE DIFFERENT SYSTEMS; THE BASELINE SYSTEM IS THE STANDARD LIA SPEECH RECOGNITION SYSTEM (SPEERAL) WHITHOUT HELP, USING ONLY ONE DECODING PASS; DDA+IT = DDA + IMPERFECT TRANSCRIPT; DDA+PT = DDA + EXACT TRANSCRIPT

of all available information: we obtain a better automatic transcription and a better alignment due to the decoder quality. These two criteria increase the size of corpus in only one pass. The search of segments is fast and incremental. In addition experiments with more than one pass show that the system converges to the best potential solution during the first pass.

D. Results on RTBF in real condition

The RTBF is a Belgium broadcast news radio. We used 200 hours of speech signal with associated prompts. Prompts are grouped by month and are imperfect. This data allows us to measure the quality of our approach. A language model is estimated on all prompts (about 2400000 words) and merged with a generic language model. Baseline results are presented in table VII from one pass decoding associated with an *a-posteriori* alignment with the text island spotting algorithm: we obtain about 30 hours of exact annotated speech. In a second time we use the transcript island driven decoding. Then the quantity of aligned data is compared. Results show that with the transcript island driven decoding, 38% of additionnal words are aligned to the prompts: we have 50 hours of exact annotated speech. Moreover, without DDA, spotted segments are similar: the robustness of the spotting island algorithm is retained. These results show significant improvements in the quantity of usable data. The DDA allows us to correct on the fly the ASR system: the increase of data quantity should result in a WER improvement. This experiment brings expected results: a larger corpus with better quality.

System	Baseline	Driven Decoding
# Hours	200	200
# segments	50370	50370
# decoded words	2 497 125	2 515 503
# aligned segments	11158 (22%)	11487 (23%)
# aligned words	380042 (15%: about 30 hours)	615481 (25%: about 50 hours)

TABLE VII

THIS TABLE SHOW THE NUMBER OF MATCHING WORDS BETWEEN THE PROMPTS AND THE ASR OUTPUT. THE FIRST COLUMN SHOW THE BASELINE SYSTEM AND THE SECOND COLUMN SHOW THE TRANSCRIPT ISLAND DRIVEN DECODING METHOD

IV. CONCLUSION AND FUTURE WORK

We have proposed a method to produce a high quality corpus with a set of recorded audio which have inaccurate prompts and deprived of timecodes. The selection of matching segments showed very good results, while the DDA dramatically improved the decoder quality.

The method was developed in two parts :

- The text island spotting algorithm allows us to synchronize on demand segments with the automatic speech recognition system. The spotting offers a very good F-measure (section III-B.3) in real time with large databases of prompts. This allows us to insert the spotting algorithm directly into the decoding process.
- Once the segments are synchronized, DDA is able to correct both prompt and the ouput decoder. As DDA is synchronized only with good segments, the decoding process time is decreased.

The combination of the two proposed algorithms allows us to extract only aligned segments and to correct inaccurate words. This method rapidly produces high quality corpus. Moreover, experiments on the broadcast news RTBF show that the method produces a larger corpus than other *a-posteriori* approaches. We plan to measure the gains of this approach using the RTBF database for estimate acoustic models.

REFERENCES

[1] P. Cardinal, G. Boulianne, and M. Comeau, "Segmentation of recordings based on partial transcriptions," *INTERSPEECH 2005*, 2005.

[2] P. Placeway and J.Lafferty, "Cheating with imperfect transcripts," *Spoken Language ICSLP 96. Proceedings*, pp. 2115–2118 vol.4.

[3] M. J. Witbrock and A. G. Hauptmann, "Improving acoustic models by watching television," 1998.

[4] P. J. Moreno, C. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," *International Conference on Spoken Language Processing*, 1998.

[5] M. Mohri, "Edit-distance of weighted automata," *CIAA*, 2002.

[6] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," *AAAI Workshop on Knowledge Discovery in Databases, KDD-94*, 1994.

[7] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer, "The 1996 hub-4 sphinx-3 system," *Proceedings of the 1997 ARPA Speech Recognition Workshop, pp. 85-89*, Feb. 1997.

[8] A. Haubold and J. Kender, "Alignment of speech to highly imperfect text transcriptions," *Multimedia and Expo, 2007 IEEE International Conference on*, 2007.

[9] H. Chih-wei, "Automatic closed caption alignment based on speech recognition transcripts," 2003.

[10] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic models training," *Computer Speech and Language*, vol. 16, pp. 115–229, 2002.

[11] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol. 147*, 1981.

[12] D. A. Hull., "Stemming algorithms: A case study for detailed evaluation."

[13] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Information Processing and Management*, 1988.

[14] H. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," *Proceedings of ICSLP*, 2004.

[15] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," *ICASSP*, 2004.

[16] B. Lecouteux, G. Linarès, J. Bonastre, and P. Nocéra, "Imperfect transcript driven speech recognition," in *Interspeech'06-ICSLP*, Pittburgh, Pensylvania, USA, 2006.

[17] P. Nocéra, G. Linarès, and D. Massonié, "Phoneme lattice based a* search algorithm for speech recognition," *Text, Speech and Dialogue : 5th International Conference, TSD 2002, Brno, Czech Republic*, 2002.

[18] G. Linarès, P. Nocéra, D. Matrouf, F. Béchet, D. Massonié, and C. Fredouille, "Le système de transcription du lia pour ester-2005," 2005.

[19] B. Lecouteux, G. Linarès, F. Beaugendre, and P. Nocéra, "Text island spotting in large speech databases," *Interspeech*, 2007.

V. BIOGRAPHIES

B enjamin Lecouteux obtained a Master in 2005 in the field of automatic speech recognition. He joined the LIA in 2005 as Ph.D. student, in the Speech Processing group. His current research interest are Automatic Speech Recognition and System Combination.

G eorges Linares obtained a Ph.D. in 1998 in the field of Neural networks for acoustic processing. He joined the LIA in 1999 as associate professor, in the Speech Processing group. He participates activelly to the development of the speech recognition system of the LIA (SPEERAL), specially for acoustic modeling and the real time engine. His current research interest are acoustic modeling for Automatic Speech Recognition, decoding strategies, and indexing audiovisual databases.