Text island spotting in large speech databases

B. Lecouteux[1], *G. Linarès*[1], *Frédéric Beaugendre*[2], *Pascal Nocera*[1]

[1]Laboratoire Informatique d'Avignon (LIA)
University of Avignon, France
[2]Voice-Insight
Brussels, Belgium.

benjamin.lecouteux,georges.linares,pascal.nocera@univ-avignon.fr,
frederic.beaugendre@voiceinsight.com

10.21437/Interspeech.2007-404

Abstract

This paper addresses the problem of using journalist prompts or closed captions to build corpora for training speech recognition systems. Generally, these text documents are imperfect transcripts which suffer from the lack of timestamps. We propose a method combining a driven decoding algorithm and a fast-match process allowing to spot text-segments. This method is evaluated both on the French ESTER ([1]) corpus and on a large database composed of records from the *Radio Television Belge Francophone (RTBF)* associated to real prompts. Results show very good performance in terms of spotting; we observed a F-measure of about 98% on spotting the real text island provided by the RTBF corpus. Moreover, the decoding driven by the imperfect transcript island outperforms significantly the baseline system.

Index Terms: speech recognition, closed captioning, corpus building

1. Introduction

HMM based automatic speech recognition (ASR) systems require very large amounts of annotated training data, especially on LVCSR tasks where acoustic models are composed of several millions of parameters. Moreover, discriminative methods such as MMIE (Maximum Mutual Information Estimation) or MPE (Minimum Phone Error) are now widely integrated in speech recognition systems. This kind of methods brings significant performance gains as far as sufficient amount of annotated data are available for model estimate. Thus, recent state-of-the-art ASR systems are typically based on corpora composed of more than 1000 hours of training data [2]. Such a constraint constitutes a major limit in applying speech technology to low resourced languages and might increase dramatically the cost of LVCSR systems construction.

Nevertheless, low cost transcripts could be available when text sources are associated to speech materials. Especially, the prompting or the captioning of broadcast news could provide approximated transcripts of large speech recordings.

Unfortunately, two major difficulties limit the use of prompts or closed captions as training corpora.

The first is that they are generally imperfect transcripts, as speakers may not strictly follow the prompter. [3] reports about 10% WER (Word Error Rate) in TV closed captions; some authors proposed methods for taking advantage of these imperfect transcripts for acoustic models training or adaptation. [4] and [5] have shown that lightly supervised adaptation to closed-captions could improve significantly acoustic models precision. Previously, we proposed a *Driven Decoding Algorithm* (DDA) which is able to simultaneously align and correct the imperfect transcripts [6].

The second problem is that closed captions usually suffer from both time lag due to the lack of post-synchronization on the audio material and that numerous speech segments of variable lengths are not prompted; this particularly occurs when speakers alternate between read talks (following a prompter for example) and more spontaneous comments or interviews (not transcribed).

This problem has been tackled by some authors under the more general topic of text-to-speech alignment ([7],[8]). Most of the proposed methods rely on DTW algorithm where the available word utterance is synchronized to the outputs of a speech recognition system. Due to the intrinsic complexity of the DTW algorithm, these methods are efficient if the two following constraints are satisfied : (1) the size of speech segments to be processed must match the computer capabilities. This means that speech streams must be segmented into trackable parts into which the selected text map onto the corresponding speech signal; (2) a typical way to improve DTW efficiency consists in pruning paths which correspond to high temporal distortions between signal and transcript. These heuristics fail when large transcripts or speech segment are missed.

Other related works propose to solve this problem of pre-synchronization using multi-pass strategies. Reported results show a high efficiency in terms of alignment scores, at the cost of a relatively heavy iterative process.

In this paper, we address the problem of spotting of transcript islands, in the general framework of driven decoding methods.

The next section describes the proposed approach. We first present the principle of the driven decoding algorithm (DDA). Then, we focus on our fast-match algorithm for transcripts spotting. We show how this method deals with missing transcripts and we describe it's integration into the DDA global scheme

Section 3 reports and comments experiments. We present the RTBF and ESTER databases on which the experiments are carried out; then, the proposed method is evaluated both in terms of transcript spotting and error corrections.

Finally, section 4 concludes and suggests new applications of the proposed method for spotting and correcting transcript island.

2. Spotting transcript Islands

2.1. Imperfect transcript Driven decoding

A driven decoding algorithm aims to align and correct imperfect transcripts by using a speech recognition engine. The algorithm proceeds in two steps. Firstly, the provided transcripts

h_p and the current hypothesis h_c are synchronized by a DTW algorithm. When the anchorage points are found, an hypothesis-to-transcript matching score is computed and used for linguistic rescoring. This algorithm improves dramatically the system performance by taking advantage of the availability of the approximative transcripts. Moreover, our evaluation on the ESTER corpus has shown that driven decoding improves significantly the quality of the initial transcripts. Our previous experiments have shown that corrected transcripts outperforms the initial ones by about 25% relative WER.

Nevertheless, the lack of significant parts of transcripts causes failures in the search of anchorage points. Therefore, the algorithm is not really relevant for the spotting of text island in large speech records. A second experimental outcome shows that when the decoder is driven by irrelevant transcripts, the search algorithm slows down by a factor from 2 to 3, depending of the system configuration, while the WER remains unchanged.

In order to apply DDA to transcript island spotting we add a fast-match process which aims to find on-the-fly the transcript islands which are relevant to the current state of the search algorithm. This method is described in the next section.

2.2. Fast-match to transcript island

The goal of our method is to take advantage of imperfect transcripts when they are available, while no timing information is available for the localization of transcript island.

On the other hand, DDA is able to integrate, in the search algorithm, some information related to the prompts. The aim is to combine to DDA an island detector which would be able to decide, at each node of the search graph, when the recognizer is crossing a transcript island. As the search graph is developed dynamically, this must be achieved by an on-the-fly spotting process.

The principle of the proposed method is close to approaches used in the field of information retrieval. In our case, the hypothesis is a query which may be answered by one of the transcript island. Typically, search engines try to find the most relevant documents by comparing the query to the indexed collection of stored documents. Most of the algorithms consist in building a set of ranked document lists. Here, we follow a similar scheme, while focusing on the efficiency of the algorithm.

The lexicon is represented by a lexical space L_s where each dimension is associated to a word. All documents, including the hypothesis itself, are represented in this lexical space by word-frequency vectors. The coefficients of these vectors represent the frequencies of words in the document.

As the current hypothesis is developed, a set of word clusters C_i is built and updated. These clusters result from the intersection of h_c and the transcript island I_i. For each new word added to the hypothesis h_c, transcript islands are considered as candidates for guiding the search. This competition is arbitrated by a matching score W_i which is computed as follow :

$$W_i(h_c) = \frac{\mid C_i(t) \mid}{\mid h_c(t) \mid} * \sum_{i=0}^{n} Idf(w_i)$$

where $\mid C_i(t) \mid$ and $\mid h_c(t) \mid$ are the cardinality of respectively the cluster C_i and the current hypothesis h_c. $Idf(w)$ represents the classical measure of the relative word frequency :

$$idf(w) = \frac{1}{frequency_w}$$

Therefore, this matching score represents a level of similarity between the hypothesis and the transcript island considered. This measure takes into account the semantic weight of the word, which depends from it's relative frequency in the document.

If the higher weight is greater than an *a priori* fixed threshold, the algorithm considers that it is on a transcript island and the search algorithm is driven by the corresponding word utterance.

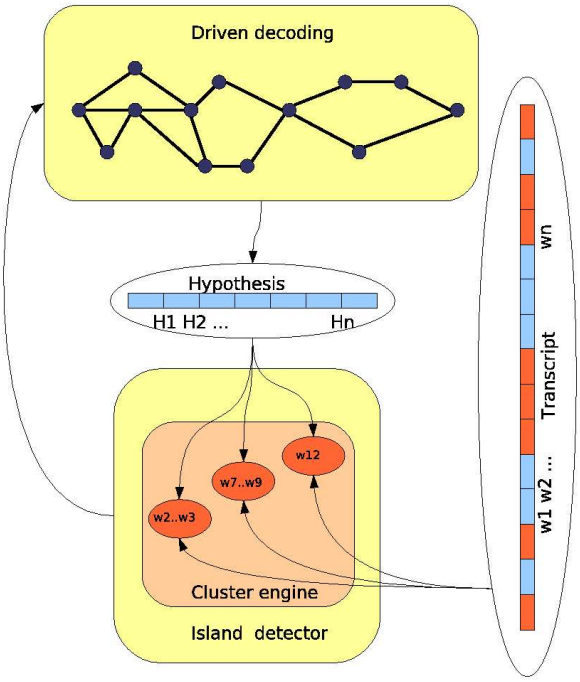

Figure 1: Scheme of principle of spotter integration to Driven decoding. Transcript island are detected by the spotter which compute an hypothesis-to-island matching score. According to it, the spotter set the decoder in driven decoding mode. Then, linguistic probabilities are rescored according to the guiding word-utterance.

3. Experiments

3.1. Experimental framework

3.1.1. The LIA broadcast news system

Experiments are carried out by using the LIA broadcast news (BN) system which was used in the ESTER evaluation campaign. This system relies on the HMM-based decoder developed at the LIA (Speeral, [9]). The segmentation tools come from the Alize toolbox ([10]).

Speeral is an asynchronous decoder operating on a phoneme lattice; acoustic models are HMM-based, context-dependent with cross word triphones. These models are composed of about 230000 gaussian components and 3600 tied states estimated on the ESTER materials (about 80 hours of annotated speech).

The language models are classical trigrams estimated on about 200M words from the French newspaper *Le Monde* and from the ESTER broadcast news corpus (about 1M words).

The full BN system runs two passes. The first one provides intermediate transcripts which are used for MLLR adaptation.

The first pass takes about 3xRT and the second one about 5xRT on a standard desktop computer (with an Opteron at 2Ghz). In these experiments, only one pass is performed, according to our initial goal of a fast word-sequence spotting. Moreover, experiments show that driving the decoding during the first pass allows to reach performance close to the one obtained by performing 2 passes [11].

3.1.2. Evaluation corpora

Two corpora are used for evaluation. The first one is based on the corpus provided for the ESTER evaluation campaign; we modified transcripts in order to simulate the fragmented closed captions which correspond to the targeted task. The second corpus is a real-world database composed of recording of RTBF broadcasts in which some segments were prompted.

3.1.3. ESTER corpus

The first test is based on the corpus provided for the ESTER evaluation campaign. The ESTER corpus consists of French radio broadcasts of the Radio-France group. It has been designed to evaluate transcription systems; therefore, it contains high quality transcripts, as opposed to the RTBF database.

ESTER broadcasts are essentially composed of news; nevertheless, some segments are recorded in more difficult acoustic conditions, such as ad-hoc interviews, talk of non-native speakers, on-the-fly translations... We use a test corpus composed by 3 hours from the ESTER development set. Imperfect transcripts are generated by adding errors manually in the initial transcripts, while ensuring a correct journalistic form in order to respect the traditional style of a radio broadcast: 10% WER is introduced in 2 hours of transcripts, and 20% WER in the last one. Moreover, about 50% of transcript segments are removed in order to simulate the lack of transcript parts.

3.1.4. RTBF corpus

The second corpus has been collected in the framework of the AIDAR project ([12]). It contains about 1000 hours of radio programs from the *Radio Television Belge*, in French and mostly recorded under clean conditions. Those programs mainly consist of news whose topics and linguistic style are rather close to the ones of the ESTER corpus. Among those 1000 hours, prompts (provided in XML files) are available for about 60 hours of news program and the proportion of actual speech exceeds 300 hours. These prompts were effectively used by the journalists. No further refinement on transcripts was done after recording and no timestamp is available in order to precisely locate the speech segments corresponding to the provided transcripts. In order to evaluate our algorithm on this database, we have manually annotated the timestamps related to prompted segments for 11 hours of the corpus.

3.2. Results

These experiments aim to evaluate the performance of the proposed method for building a speech corpus of decent quality using the closed-captions associated to the speech signals. The system performance on the spotting task is evaluated in terms of precision/recall rates, on ESTER and RTBF corpora. F-measures are also reported.

We first test alignment and correction performed on the ESTER corpus, using the exact and the degraded transcripts. Then, we consider the WER of corrected transcripts, according to the quality of text sources involved in the decoding process. Finally, we evaluate our method on the RTBF corpus based of real close captioning.

3.2.1. Spotting exact transcript island on ESTER database

Here, we evaluate the performance of our spotter using the exact transcript. The removed transcripts have been chosen according to the reference speech segmentation, by deleting randomly 50% of the speech segments. The average duration of remaining segments is about 6 minutes.

Results are reported in Table 1. We can see that, in these simulated conditions, spotting performance is good; more than 95.3% of segments have been found, with a precision of about 96.7%. Results seem relatively independent from the performance of the ASR systems which are varying, in this test, from 27.2% (RFI show) to 22.6 (France Inter show).%.

Radio station	Precision	Recall	F-measure	Seg. number
INTER	90.9%	98.89%	94.8%	478
INFO	93.7%	92.9%	91.5%	468
RFI	98.9%	97.8%	98.4%	812
Mean	95.3%	97.3%	95.5%	1758

Table 1: Transcript island spotting on the ESTER database. Experiments are performed on 3 hours of the development set, from France Inter radio (INTER), France Info (INFO) and Radio France International (RFI). The targets of spotting are **exact** transcript island.

3.2.2. Spotting imperfect transcripts on ESTER database

The second experiment aims at evaluating how errors in transcript impact the system performance. Results are reported in Table 2. We observe that the obtained precision and recall rates are very close to the ones obtained on perfect transcripts. Nevertheless, this experiment is achieved by setting a transcript word error rate of about 10%; this rate corresponds to the one reported in the literature (about 10%). A strong increase of transcripts WER should impact dramatically the spotter performance. An extreme situation would consist in submitting to the spotter a totally erroneous transcript, in which no information could allow spotting.

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	94.3%	94.8%	94.5%	1758

Table 2: Spotting on the ESTER database. Experiments are performed on 3 hours of the development set, using imperfect transcripts of about 10% Word Error Rate. As in Table 1, 50% of transcript segments have been removed for spotting evaluation.

3.2.3. Transcript island Driven decoding

DDA method allows to improve recognition rates by taking benefit of the available transcripts, even if they are not perfect. Here, we evaluate the quality of transcripts provided by the Speeral decoder guided by the transcripts themselves. Three tests have been performed and compared to the baseline system. This last one consists in a classical Speeral run, without any helpful transcript.

We first evaluates WER by using segments of exact transcripts; then, the same experiment is achieved on imperfect transcripts. Results are reported in Table 3.

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 3: Word error rates of the systems involved in the experiments; the baseline system is the standard LIA speech recognition system (Speeral) without using any text sources for helping the search, using only one decoding pass; DDA+IT consists in driven decoding by imperfect transcripts; DDA+PT is DDA search algorithm driven by the correct word utterance. Experiments are performed on 3 hours of the development set, by using imperfect transcripts of about 10% Word Error Rate. As in Table 1, 50% of text-segments have been removed for spotting evaluation.

Results show that the driven recognizer takes advantage of spotted transcripts. Of course, the correct prompts remain more efficient that imperfect ones; nevertheless, approximative transcripts bring a WER gain of about 14% relative, while exact ones allows a WER gain close to 24% relative.

3.2.4. Spotting real prompts on RTBF corpus

	Precision	Recall	F-measure	Seg. number
RTBF shows	99.28 %	97.13 %	98.41 %	501

Table 4: Precision/recall and F-measure of transcript island spotting. Experiment is performed on broadcast news shows from RTBF, by using the real journalist prompts.

Here, we test our spotting technique on the RTBF database. Experiments are performed on the 11 hours on which time stamps were manually added. The estimated WER of RTBF transcripts is lower than 5% while the base is only partially annotated, the quality of transcripts remains relatively high.

Results are quite good, and significantly better than the ones observed on the ESTER corpus. 10 from the 22 (of 30 minutes each) audio segments are fully timestamped, and the averaged F-Measure is around 98%. These better results are probably due to the fact that the prompts mapping is usually related to the global structure of the document, including speaker and speech/non-speech turns, etc. Then, transcript islands match to this natural segmentation of the document. We also can observe that well segmented transcripts are easier to spot. Moreover, the number of text segments is significantly higher in the RTBF corpus : we have about 20 annotated prompt parts per audio segment, for a total of about 400 ; this probably decreases the risk to miss an island which should be spotted.

4. Conclusion and future work

In this paper, we proposed a method dedicated to the spotting of transcript island in large speech databases. This method aims to recover the missing timestamps of closed captions, allowing to build low-cost databases for speech recognizer training. This spotter is integrated in the general scheme of driven decoding. Its role consists in dynamically detecting the transcript island, at the moment where the recognizer crosses it. Our experiments have shown that the proposed technique reaches very good results by using the real prompts provided with the RTBF database. Moreover, this method seems to be quite robust to the imperfect transcripts.

We plan now to extend our experiments to evaluate this technique on a citation spotting task, which corresponds to the particular case where only one word-utterance is spotted. This targeted task involves strong efficiency constraints, as all relevant speech segments must be decoded for spotting the targeted word-utterance. We plan to investigate a two-level method where a first very fast-match pass selects relevant areas, before re-evaluation by the proposed driven spotting technique.

Moreover, we should evaluate this algorithm in the general framework of low-cost corpus building for low-resourced languages. We actually plan to train acoustic models on a subcorpus composed of spotted and corrected RTBF prompts. We will study how the particularities of such a corpus could be advantageously integrated in the general scheme of lightly supervised training.

5. References

[1] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news," in *Proc. of the European Conf. on Speech Communication and Technology*, 2005.

[2] R. Prasad, S. Matsoukas, C.-L. Kao, J. Ma, D.-X. Xu, T. Colthurst, O. Kimball, R. Schwartz, J. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre, "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System," in *InterSpeech 2005*, Lisbon, 2005.

[3] P. Placeway and J. Lafferty, "Cheating with imperfect transcripts," *Proceedings of ICSLP*, 1996.

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

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

[6] B. Lecouteux, G. Linares, J. Bonastre, and P. Nocera, "Imperfect transcript driven speech recognition," in *InterSpeech'06*, 2006.

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

[8] P. J. Jang and A. G.Hauptmann, "Improving acoustic models with captioned multimedia speech," *IEEE International Conference on Multimedia Computing and Systems, Florence, Italy*, 1999.

[9] P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, and F. Béchet, "The LIA's French broadcast news transcription system," in *SWIM: Lectures by Masters in Speech Processing*, Maui, Hawaii, 2004.

[10] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *ICASSP'05*, Philadelphia, USA, March 2005.

[11] B. Lecouteux, G. Linares, Y. Esteve, and J. Mauclair, "System combination by driven decoding," in *ICASSP'07*, 2007.

[12] B. F. B. G. M. G. Tshibasu-Kabeya, B., "Aidar : Une architecture pour lindexation de documents audionumriques," in *VSST 2006*, Lille, France, January 2006.