Imperfect transcript driven speech recognition

Benjamin Lecouteux, Georges Linarès, Pascal Nocéra, Jean-François Bonastre

Laboratoire d'Informatique d'Avignon
339, chemin des Meinajaries
Agroparc – B.P. 1228
F-84911 Avignon Cedex 9

{benjamin.lecouteux,georges.linares,pascal.nocera,jean-francois.bonastre}@univ-avignon.fr

Abstract

In many cases, textual information can be associated with speech signals such as movie subtitles, theater scenarios, broadcast news summaries etc. This information could be considered as approximated transcripts and corresponds rarely to the exact word utterances. The goal of this work is to use this kind of information to improve the performance of an automatic speech recognition (ASR) system. Multiple applications are possible: to follow a play with closed caption aligned to the voice signal (while respecting to performer variations) to help deaf people, to watch a movie in another language using aligned and corrected closed captions, etc. We propose in this paper a method combining a linguistic analysis of the imperfect transcripts and a dynamic synchronization of these transcripts inside the search algorithm.

The proposed technique is based on language model adaptation and on-line synchronization of the search algorithm. Experiments are carried out on an extract of the ESTER evaluation campaign [4] database, using the LIA Broadcast News system. The results show that the transcript-driven system outperforms significantly both the original recognizer and the imperfect transcript itself.

Index Terms: speech recognition, approximated transcripts.

1. Introduction

In many cases, external information sources associated with speech signal are available: for instance movie subtitles or broadcast news with summary, theater scenario, journalist prompt. This information could be used in order to improve the performance of an automatic speech recognition (ASR) system.

However, when the available transcript differs from the correct one, speech-to-text alignment methods fail and large speech segments may be removed. In many cases, an automatic speech recognizer could provide useful but probably erroneous transcripts. Nevertheless, the speech recognition system can take benefit of the provided transcripts.

The use of imperfect transcripts for audio decoding has already been studied within the framework of audio/video automatic indexing [2] or for acoustic model re-estimation with non transcribed data [5][6]. These works are frequently motivated by the availability of large corpora which are partially or poorly transcribed.

This task is presented in literature as a speech-to-text synchronization problem, or as an imperfect transcript correction problem. In both cases the main difficulty resides in the low quality of the provided transcripts: P. Placeway [11] measures a WER of 10 to 20% between movie subtitles and the exact transcript. These divergences increase considerably the alignment difficulty.

In this paper, we first present issues involved in imperfect text alignment on a word stream. Then, we describe our method for speech recognition driven by imperfect transcripts. We especially investigate the integration of this source of information to an asynchronous stack decoder. Finally, the undertaken experiments and the results obtained are detailed and some conclusions and future prospects are presented.

2. Speech-to-text alignment

2.1. Alignment with exact transcripts

Speech-to-text alignment has been widely studied during last decades. Several authors proposed methods based on a forced Viterbi decoding, sometimes driven by heuristics reducing the algorithm complexity.

In [8], P.J. Moreno proposes 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 well synchronized areas, called *small confidence island*. Initially, a language model is estimated on the exact transcript. Synchronized areas are isolated from the extracted segments with a high matching between the *a priori* transcript and the automatic transcript. Documents are then segmented according to these small confidence islands; on each part, a specific language model is estimated. The algorithm is launched recursively on unaligned parts until reaching the convergence point. This method, restricted to exact transcripts, obtains excellent results: 99% of the words are correctly aligned.

2.2. Alignment with imperfect transcripts

Alignment on partially incorrect transcripts has been mainly used to deal with corpora for which only low quality transcripts are available. The goal is to improve the database usability, generally by extracting well-transcribed speech segments.

In [6], L. Lamel evaluates a method for acoustic model training on low quality transcribed databases. This technique consists in decoding the training database automatically and comparing the resulting hypotheses to the imperfect transcripts. Matching segments are then used for acoustic model re-estimation. This process is iteratively performed until convergence. In this paper, we want not only to find well-transcribed segments but also to improve the

transcripts on segments where no correct transcripts are available. This issue was tackled by P. Placeway [11] to take advantage of subtitles with a synchronous decoder (Sphinx-3, [10]). Their experiments are based on an English broadcast news database. They propose to estimate a language model on the subtitles and align them to the audio stream.

In order to combine information from closed captions and language models, P. Placeway [11] proposes to interpolate a generic model and a model estimated on the subtitles. This interpolated model is then used by the ASR. Their experiments are carried out on 9.7% WER subtitles. Using a beam-search decoder, this technique improves the performance of 15% relative WER (Word Error Rate) compared to the initial decoding (from 55.8% to 47.2%). Moreover, they propose to integrate a time warping algorithm in addition to the model interpolation technique: as the decoder progresses and selects a small list of candidate words, the words corresponding to the imperfect transcript are favored in the hypothesis beam. This method brings a 37% relative gain in WER compared to initial decoding (while decreasing the absolute WER to 35%). The final result remains however definitely lower than the quality of the imperfect transcripts provided to the system (9.7% of WER).

3. Dealing with imperfect transcripts

Our objective is to exploit imperfect transcripts with an asynchronous decoder based on the A* algorithm, within the framework of a broadcast news (BN) system. First, we present the characteristics of this kind of decoder. Then we show how the information from imperfect transcripts can be integrated in the decoding process.

Two methods exploiting the imperfect transcripts are proposed. The first one consists in combining a generic language model and a language model estimated on the imperfect transcript. The second one integrates a Viterbi-based alignment algorithm within the A* algorithm, by on-demand synchronization and estimate function rescoring.

3.1. Language model adaptation

In our approach, linguistic variability can be reduced thanks to the contribution of an exact or imperfect transcript. A profit can be obtained by reducing overall linguistic space. This can be achieved by estimating a language model on the transcript itself. However, such a language model would be probably too specific when the speaker deviates from the original transcript. Therefore, this model is interpolated with a generic language model.

3.2. Anatomy of SPEERAL decoder

LIA has developed a large vocabulary continuous speech recognition system named SPEERAL [9]. This decoder is derived from the A* search algorithm operating on a phone lattice. The exploration of the graph is supervised by an estimate function $F(h_n)$ which evaluates the probability of the hypothesis h_n crossing the node n:

$$F(h_n) = g(h_n) + p(h_n) \quad (1)$$

where $g(h_n)$ is the probability of the current hypothesis which results from the partial exploration of the search graph (from the starting point to the current node n); $p(h_n)$ is the probe which estimates the probability of the best hypothesis from the current node n to the ending node.

In SPEERAL, the probe p combines the acoustic probability and a linguistic look-ahead score [3]. The acoustic term is computed by an acoustic decoding carried out by the Viterbi-back algorithm operating on a phone lattice.

The graph exploration is based on the function of estimate $F()$. Indeed, the stack of hypotheses is ordered on each node according to $F()$. The best paths are then explored firstly. This deep search refines the evaluation of the current hypothesis. Low-probability paths are cutted-off, leading to search backtrack. In such situations, search is desynchronized from the audio stream.

In order to be able to take information resulting from the imperfect transcript into account, the $F()$ function is modified to influence the score of the current hypothesis. In fact, this mechanism drives the search by dynamically rescoring $g(h_n)$ according to the alignment scores. This algorithm consists in two parts: synchronization of the transcript and integration of the synchronized transcript inside the A^* evaluation function.

3.3. Audio stream synchronization to the imperfect transcript

The speech recognition system generates hypotheses as it forwards in the phoneme lattice. The best hypotheses at time t are extended according to the current hypothesis probability and the probe results. In order to locate an anchorage point into the transcript, each evaluated word is aligned to the reference word stream using an algorithm of temporal alignment (Dynamic Time Warping [1]). A partial hypothesis is built by collecting the current word and its history from the path found during the search process. The best hypothesis-to-reference matching provides a synchronization point which will be used for hypothesis rescoring.

Considering the low complexity of the word alignment, this on-demand synchronization process requires low CPU resources. Moreover, the additional alignment cost may be balanced by the speed-up provided on well-transcribed sections.

Figure 1 illustrates the dynamic synchronization of the search driven by alignment on an imperfect transcript.

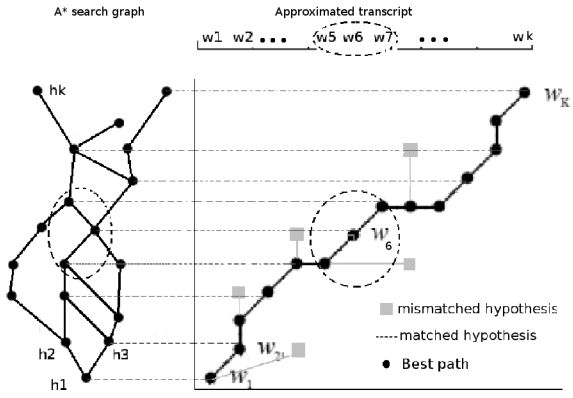

Figure 1: Synchronization of the search beams with the imperfect transcript by the DTW algorithm during asynchronous decoding

3.4. Weighting of the current hypothesis according to alignment

Once the hypothesis is synchronized with the transcript, the algorithm estimates the matching transcript-to-hypothesis score (noted

α). This score is based on the number of words in the short-term history which are correctly aligned with the transcript: only three values are used, corresponding to a full alignment of the current trigram, a full alignment of the current bigram and an alignment of one word only. α is maximum when the trigram is aligned and decreases according to the misalignments of the history. Values of α are determined empirically using a development corpus : $0.9; 0.7; 0.3$. Then, linguistic probabilities are modified using the following rescoring rule:

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

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

4. Experiments

4.1. Experimental context

The experiments are carried out with the "Broadcast news" system developed by the LIA for the French evaluation campaign ESTER [7].

4.2. Corpora and imperfect transcript

The system is assessed on 3 hours of radio broadcast (France Inter 1, France Inter 2, France Info) extracted from the ESTER development corpus. Imperfect transcripts are generated by adding errors manually in the initial transcript, while ensuring a correct journalistic form in order to respect the traditional style of a radio broadcast: 10% WER are introduced in 2 show transcripts, and 20% WER are introduced in the last show transcript.

4.3. Language model interpolation

Some preliminary experiments are undertaken in order to identify the potential benefits of the proposed methods. First, a language model is estimated on the exact transcript. Then, this model is combined with a generic language model (65000 words learned on the newspaper Le Monde). The objective is to measure the real effect of the suggested techniques on the decoder recognition performance[1]. Table 1 shows the results of the interpolation of a language model trained on the exact transcript with a generic language model. For comparison, a baseline decoding is performed by using a generic language model, we obtain a WER of 22.7%.

	WER
FrInter 1: ML-G alone	22.7%
FrInter 1: ML-TrEx alone	5.2%
FrInter 1: ML-G 70% + ML-TrEx 30%	13.0%
FrInter 1: ML-G 50% + ML-TrEx 50%	11.5%
FrInter 1: 30% + ML-TrEx 70%	10.8%

Table 1: Interpolation of the generic language model (ML-G) with a model trained on the exact transcript (ML-TrEx)

Then, starting from the imperfect transcript (10% WER), a language model is also generated. The experiments using this language model combined with the generic model are presented in table 2.

[1]The words not found in the ASR lexicon are extracted from the transcript and added to the language model

	WER
FrInter 1: ML-TrErr alone	16.3%
FrInter 1: ML-G 70% + ML-TrErr 30%	16.2%
FrInter 1: ML-G 50% + ML-TrErr 50%	15.4%
FrInter 1: ML-G 30% + ML-TrErr 70%	15.2%

Table 2: Interpolation of the generic language model (ML-G) with the model trained on the imperfect transcript (ML-TrErr - 10% WER)

These experiments show that a decoding based on a language model estimated on the imperfect transcripts improves significantly the WER. However, without another information source, the speech recognition system converges to errors contained in the transcript. This technique does not allow to obtain good recognition rates. Indeed, using a language model estimated on the exact transcript, the WER remains under 15% despite the availability of the perfect transcripts. So we need to use more efficiently transcript information.

4.4. Experiments with model interpolation and dynamic synchronization

This section presents experiments using the decoding strategy based on both dynamic synchronization and linguistic rescoring.

The experiments combining the interpolation of the language models with an alignment on the exact transcript are presented in table 3.

	WER
FrInter 1: ML-G alone + alignment TrEx	3.7%
FrInter 1: ML-TrEx alone + alTrEx	3.7%
FrInter 1: ML-G70%+ML-TrEx30%+alTrEx	4.9%
FrInter 1: ML-G50%+ML-TrEx50%+alTrEx	3.5%
FrInter 1: ML-G30%+ML-TrEx70%+alTrEx	3.7%

Table 3: Interpolation of the generic language model (ML-G) with the model trained on the exact transcript (ML-TrEx) and alignment to the exact transcript (alTrEx)

We obtain in this case a WER of 3.5%. This level of error can be considered as minimal for a method re-estimating the concurrent hypothesis without modifying the content of the hypothesis stack.

Table 4 shows the experiments replacing the exact transcript with the imperfect transcript.

	WER
FrInter 1: ML-TrErr + alTrErr	9.9%
FrInter 1: ML-G + alTrErr	7.7%
FrInter 1: ML-G70%+ML-TrEr30%+alTrErr	7.2%
FrInter 1: ML-G50%+ML-TrEr50%+alTrErr	7.4%
FrInter 1: ML-G30%+ML-TrEr70%+alTrErr	8.6%

Table 4: Interpolation of the generic language model (ML-G) with the model trained on the imperfect transcript (ML-TrErr - 10% WER) and alignment to imperfect transcript (alTrErr - 10% WER)

Although this approach removes some of the limits observed in the model combination, potential sources of error remain. In particular, heuristics are used in the decoder to reduce the search

space and to accelerate decoding. In usual conditions, pruning should introduce only few errors; however, when the acoustic context is of low quality, the best hypothesis can be excluded from the stack of available hypotheses. This occurs more frequently in real time configurations of the system, for which pruning is more strict. In this case, a strategy based on the promotion of the synchronized hypothesis does not allow error recovering. One can quantify these phenomena by using the recognition engine to perform a text-to-speech forced alignment.

The best result is obtained by combining the generic language model (with a 70% weight) and the model estimated on the imperfect transcript (with 30% weight) and by carrying out an alignment on the latter. Alignment reduces WER down to 7.2%. It makes possible to bring a temporal information which is poorly taken into account by the language model. The use of a DTW alignment associated with the interpolation of the models shows a new gain.

In order to validate these results, we test the best configuration of the system on a larger corpus. Two hours are processed using the same evaluation protocol described in the last sections. Results are reported in table 5.

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 5: WER obtained by the baseline system (*Baseline*), the original transcript (*Transcript*), the Transcript Driven System (*TDS*)

We observe that the gain in performance seems to be relatively independent from the quality of the initial transcript. Nevertheless, the best improvement is obtained on the "France Info" show, which is poorly transcribed (20% WER). This result suggests a good robustness of the proposed method.

These experiments show that imperfect information can be favorably exploited during the decoding process.

5. Conclusion

A well-known advantage of A^* algorithm lies in the possibility to incorporate various information source into the recognition process. Nevertheless, it is an asynchronous algorithm and its application to alignment tasks can be difficult. We have proposed a on-demand synchronization which allows to combine asynchronous recognition and text-to-speech alignment. The system takes advantage of the approximate transcript as long as it allows a gain and switches in free-recognition mode when the acoustic observations do not match the suggested transcript.

We evaluated our approach step-by-step. For each phase we estimated the system performance and we compared the results with the maximum performance bound allowed by the proposed approach.

The first evaluated technique consists in extracting linguistic information from the script by estimating a language model on the imperfect transcripts. Our experiments showed that the interpolation of this model with the generic language model leads to significant improvements. Nevertheless this method does not allow to go further than the quality of the imperfect transcript. This drawback may limit the interest of that kind of approach.

The second method presented in this paper consists in driving the search algorithm with the provided imperfect transcripts. The method relies on a synchronization between the current search hypothesis and the provided information.

This method allows a significant gain in terms of WER even if the quality of the provided transcripts is low. The obtained relative WER improvement is between 28% and 40%.

Moreover, we observe that the modified algorithm improves slightly the decoding speed, in spite of the additional computational cost due to the search synchronization. This profit in terms of execution time is due to the earlier exploration of the best paths on well transcribed sections. However, search efficiency can be improved by introducing heuristics in the probe itself.

Although these first results show the interest of an alignment on an approached transcript, these experiments were carried out under controlled conditions: a relative low level of noise, transcript closed to the exact transcript and BN speech style and topics. We plan to use our methods under more difficult conditions, for example on automatic subtitling of movies or theater plays.

6. References

[1] 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.

[2] Huang Chih-wei. Automatic closed caption alignment based on speech recognition transcripts. 2003.

[3] G. Linarès D. Massonié, P. Nocéra. Scalable language model look-ahead for lvcsr. *InterSpeech'05, Lisboa, Portugal*, 2005.

[4] 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. Proc. of the european conf. on speech communication and technology. 2005.

[5] Photina Jaeyung Jang and Alexander G.Hauptmann. Improving acoustic models with captioned multimedia speech. *IEEE International Conference on Multimedia Computing and Systems, Florence, Italy*, 1999.

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

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

[8] Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. A recursive algorithm for the forced alignment of very long audio segments. *International Conference on Spoken Language Processing*, 1998.

[9] Pascal Nocera, Georges Linares, and Dominique Massonié. Phoneme lattice based a* search algorithm for speech recognition. *Text, Speech and Dialogue : 5th International Conference, TSD 2002, Brno, Czech Republic*, 2002.

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

[11] Paul Placeway and John Lafferty. Cheating with imperfect transcripts. *Proceedings of ICSLP*, 1996.