Transcriber driving strategies for transcription aid system

Grégory Senay, Georges Linarès, Benjamin Lecouteux, Stanislas Oger, Thierry Michel

LIA, University of Avignon, France
Xtensive Technologies, Strasbourg, France
{gregory.senay,georges.linares,benjamin.lecouteux,stanislas.oger}@univ-avignon.fr; thierry.michel@xtensive.com

Abstract

Speech recognition technology suffers from a lack of robustness which limits its usability for fully automated speech-to-text transcription, and manual correction is generally required to obtain perfect transcripts. In this paper, we propose a general scheme for semi-automatic transcription, in which the system and the transcriptionist contribute jointly to the speech transcription. The proposed system relies on the editing of confusion networks and on reactive decoding, the latter one being supposed to take benefits from the manual correction and improve the error rates. In order to reduce the correction time, we evaluate various strategies aiming to guide the transcriptionist towards the critical areas of transcripts. These strategies are based on graph density-based criterion and two semantic consistency criterion; using a corpus-based method and a web-search engine. They allow to indicate to the user the areas which present severe lacks of understandability. We evaluate these driving strategies by simulating the correction process of French broadcast news transcriptions. Results show that interactive decoding improves the correction act efficiency with all driving strategies and semantic information must be integrated into the interactive decoding process.

1. Introduction

The performances of Automatic Speech Recognition (*ASR*) systems are still strongly dependent on the acoustic and linguistic contexts. On well-defined domains, systems are relatively accurate, but performance decreases dramatically when the running conditions differ significantly from those in the training data. This lack of robustness limits the integration of speech technology in real-word applications; on speech mining or retrieval tasks, the errors on meaningful words lead to incorrect content analysis.

Considering the limits of state-of-the-art *ASR* systems, perfect automatic speech recognition remains a long-term perspective. On the other hand, the cost of manual annotation is high and *ASR* systems could provide a helpful contribution to this task. This idea has been integrated into some recent academic and industrial projects. Most of the proposed approaches consist in using the one-best hypothesis provided by a recognition engine to speed up the annotation work (Bazillon et al., 2008). However, *ASR* systems could provide supplementary information through the decoding stage, and this information could be used for reducing, as much as possible, the correction cost. This additional information could be held by the alternative hypotheses that were evaluated during the decoding process (Nanjo et al., 2006), but could also be extracted by automatically checking the system outputs. Reciprocally, the information provided by the corrector could be used by the *ASR* system to improve its own performance.

Starting from these ideas, we propose an interactive decoding strategy whereby human and computer contribute jointly to the speech transcription. In this correction scenario, a fast decoding pass follows each corrective action, integrating the corrections as templates of the searched transcription. In such an interactive decoding scheme, the order the corrections are performed may impact significantly the efficiency of the post-decoding pass. We propose various methods that aims to drive the transcriptionists toward such critical areas.

We first evaluate a criterion based on the search-graph density. Then, considering that the loss of semantic content impacts dramatically on the transcript relevance, we propose to use a semantic consistency criterion in order to guide the human corrector on critically-deficient parts of the transcripts.

2. Interactive decoding

2.1. Principle

The use of the one-best hypothesis only provides relatively poor information on the full recognition process. In fact, the ASR system evaluates a lot of alternative paths. One of the major difficulties in offering these alternatives as choices to the transcriber stems from the large number of competing hypotheses that may differ only by a few words (J.Ogata and M.Goto, 2005). Different alternatives could be proposed to the transcriber to improve this efficiency. As proposed in (D. Falavigna, 2002), we use confusion networks (CN), which are significantly more *readable* than lattices for transcript editing. Therefore, corrective actions will consist in choosing one of the proposed alternatives in the CN that results from a first decoding pass, or in manually adding a missing term.

By selecting or adding such a word in a CN section, the user provides information that may impact positively on the word neighbourhood; the principle of interactive decoding is to take advantage of this local correction by performing a fast recognition pass, constrained by the previously corrected words.

A fast decoding pass is performed after each corrective action in the CN, driven by the actual correction and the previous ones. The goal of this driving decoding pass is to improve the transcription, in particular around the words in which the linguistic context is changed. This local modification can besides globally change the transcription segment, because that may propagate positively on the entire sentence.

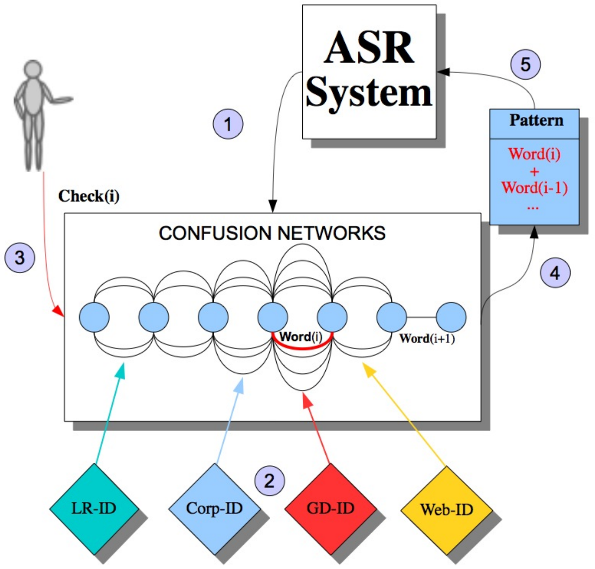

Figure 1: Interactive decoding

2.2. Pattern constrained decoding

The general scheme of the constrained decoding is to provide to the decoder a sentence pattern, wherein all the corrective actions performed by the transcriptionist (Figure 1) seem like fixed words, and the unchecked areas like wildcard, areas which must be found. Technically, interactive decoding relies on the driven decoding algorithm (DDA, (Lecouteux et al., 2006)), presented in last studies, this algorithm drives the search towards imperfect or uncompleted transcripts and system combination (Lecouteux et al., 2008). We therefore propose a decoding drive by a pattern which is created from the last transcriber corrections on the CN.

Considering this incremental correction scheme of the semi-automatic transcriptions, we can assume that the order or the areas which corrections must be checked can modified appreciably the decoding efficiency. We evaluate, in the following, various methods aiming to guide the transcriber to the areas that may be more profitable to the system in terms of Word Error Rate (WER) for each corrective act.

3. Driving Strategies

We propose to use the self-diagnostic capabilities of the ASR system for driving the transcriber to the most efficient corrections. Usually, the natural correcting process consists in performing a left-right check of the transcript; in the reading direction; and in correcting it when an error is detected. Considering the reactive decoding process, a corrective act should impact positively on the word neighbourhood, as the latter contains errors. Hence, we can expect the proposed mechanism involving the system's reaction to users' actions to be more efficient if the user chooses to correct first the sections of transcript where the expected WER gain should be maximal.

3.1. Correction driven by Graph density

Numerous methods have been developed in research related to confidence measures. Nevertheless, since confidence measure aims at estimating the correctness probability of a given word or sentence, the issue here is to maximise the expected WER gain per corrective action. The lattice depth has often been used as a relevant criterion for estimating confidence scores (Kemp and Schaaf, 1997), by using the fact that the widest area of the graph search corresponds to a situation where the search algorithm is not able to clearly identify a better path, but has trouble choosing between a large number of competing hypotheses.

Starting from this idea, we evaluate a simple strategy whereby the transcriptionist is driven towards the widest section of the CN. The system automatically highlights the highest-density areas of the graph, and proposes to the transcriber to correct them first. Each corrective act produces a corresponding sentence pattern, and is followed by a pattern-driven decoding. We estimate the efficiency of this method by computing the WER gain provided by this last decoding pass.

3.2. Correction driven by semantic consistency

The corpus-based criterion estimates the semantic consistency of the transcription by using parts of a large newswire corpus. The global approach that we propose consists in finding, in the database, the newswire that is the closest to the transcription. In order to estimate the semantic consistency measure, the system searches for the most relevant paragraph which is split into smallest segments (10 relevant words on average); for each segments there semantic consistency are evaluated and the resulting scores are used for driving the transcriptionist, which is supposed to correct the lowest-scored segments first.

The similarity between transcriptions and newswires is performed by using the Cosine metric (Rijsbergen, 1979). The information retrieval algorithm searches the database for the paragraph that maximises the Cosine metric. In order to take into account only the meaningful words, both transcriptions and corpus are lemmatised and filtered by a stoplist that contains the most frequent cue words.

These experiments are evaluated on the French Gigaword corpus. French Gigaword is an archive of newswire text data that has been acquired over several years by the Linguistic Data Consortium (LDC). The corpus was collected from two distinct international sources. This first one is the *Agence France-Presse*, which provided newswires from 1994 since 2006. This first part contains more than 480000 words. The second one comes from *Associated Press French Service*, covering the same time span and containing about 180000 words. All contents are written in the newswire style: sentences are relatively short, with the average number of relevant words per sentence is about 15. Documents are structured in paragraphs of sentences. Each paragraph corresponds to a newswire, and focus on an unique topic. The baseline corpus contains 2 millions of newswires and 250 millions of sentences.

3.3. Correction driven by Web-based criterion

The goal of this strategy is to use the Web for detecting semantic inconsistencies in the sentences. The method relies on the hypothesis that the more numerous the relevant words that appear together in Web documents are, the closer they are semantically. All the cue words are excluded using a stop-list. We propose to use the Web because it has an important language/semantic coverage. The Web is used like a big database of documents, in which each document is regarded as a bag of semantically closed words.

To estimate if a word is semantically consistent in its context, and so a word is in this Web context, we propose to use the probability that the word appears in a document, knowing that the words of its context appear.

This probability is defined in Equation 1, for a word w_i and a n-order probability, with a $n-1$ left context size $\psi_i^n = w_{i-n-1}, \ldots, w_{i-1}$:

$$P_s(w_i|\psi_i^n) = \frac{WC(\psi_i^n, w_i)}{WC(\psi_i^n)} \tag{1}$$

For a string of words $w1w2w3$, the Web-based criterion $WC(w3|w1w2)$ is the number of Web documents where the three words $w1w2w3$ occur together, normalised by the number of Web documents where only $w1w2$ occur. Like in classical n-grams, when the word with its context doesn't occur in any Web document, the model backs off to the lower-order function with a discount factor, in Equation 2. The more this score tends to 1, the closer the words are semantically. When a semantic incoherence occurs, the score falls down, thus indicating a critical zone in the transcripts.

$$\hat{P}_s(w_i|\psi_i^n) = \begin{cases} P_s(w_i|\psi_i^n), \text{ if } WC(\psi_i^n, w_i) > 0 \\ \alpha \cdot P_s(w_i|\psi_i^{n-1}), \text{ else} \end{cases} \tag{2}$$

The n-order semantic confidence score of the sentence is :

$$\begin{gathered} SC(w_1 \ldots w_i) = \\ P_s(w_2|w_1) \times P_s(w_3|w_1, w_2) \times P_s(w_i|\psi_i^n) \end{gathered} \tag{3}$$

4. Experiments

All the experiments are conducted by using the LIA broadcast news system (Speeral, (Linarès et al., 2007)) in the framework of the ESTER evaluation campaign (Galliano et al., 2005). Speeral is an asynchronous decoder operating on a phoneme lattice; acoustic models are HMM-based, context-dependent with cross word triphones. 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). Since the full BN system runs 3 passes including unsupervised speaker adaptation, we use for these experiments only the first decoding pass to produce word lattices. These lattices are mapped to confusion networks by using the SRILM toolkit (Stolcke, 2002) developed by *SRI International*. Pattern driven decoding stage is performed, as previously explained, by using the DDA.

We used a fast single-pass system (running in 2 real time) that performs a 32.6% WER on this test set. Considering a realistic usage scenario where the user should quickly get the system response, reactive decoding is performed by using the real-time system. Moreover, pattern driving reduce the size of the search space and improve the system speed. The experiments are conducted on the ESTER 2005 database development set. This database is composed by 8 hours of French broadcast news from 4 different radio stations. Nevertheless, in this paper we do not focus on decoding time, even if industrial implementation should propose a comfortable mean to exploit the ASR system responses.

4.1. Protocol

In these experiments, we start from the reference segmentation that is automatically split according to speaker turns and silence areas, respecting an upper limit: a segment size of 30 seconds. Left-right correction is automatically simulated by using the alignment provided by the *Sclite* scoring tool. Each encountered error is tagged by Sclite as confusion, insertion, or deletion. The corresponding corrective action is applied to the hypothesis, according to this tag. Corrective actions are performed following the interactive decoding-based strategies previously described: left-right correction (LR-ID), graph-density checking (GD-ID), semantically-driven correction, based on the text corpus (Corp-ID), and on the Web (web-ID). These methods are compared to a baseline that corresponds to a left-right correction without reactive decoding (Human only).

We evaluate the performances in section 4.2. by computing the absolute WER gain obtained according to the number of corrective actions per segment; There are a maximum of 20 corrective actions per segment. In order to emphasise the interactive decoding, performances are evaluated on two classes of segments, where the system yields WERs below and above 40%.

4.2. Results

In Tables 1 and 2 we show the results obtained in terms of WER, for the two classes that correspond to initial WERs below and above 40% in which segments have been corrected one, three, ten and twenty times (#c). These classes represent respectively 46% and 54% of the test corpus. We can note that the reactive decoding greatly improves the WER in all the configurations compared to the simple manual correction.

# c	1	3	10	20
Human only	25.22	22.98	17.23	9.44
LR-ID	24.28	**20.82**	**11.88**	**5.26**
GD-ID	26.58	25.38	16.62	11.76
Corp-ID	**23.90**	21.15	13.93	8.51
Web-ID	24.33	21.10	12.21	7.40

Table 1: WERs according to the number of corrective actions, for initial transcriptions of WERs below 40%.

The comparison between the various strategies in the first class (Table 1 - WER $\leq$ 40%) shows that the graph-density correction (GD-ID) seems to be rather inefficient, yielding a significantly worse WER than the classical left-right correction (with interactive decoding). It remains less accurate

# c	1	3	10	20
Human only	55.91	54.05	47.81	40.14
LR-ID	54.95	49.77	37.71	**25.36**
GD-ID	57.51	53.52	44.05	36.99
Corp-ID	54.19	49.37	39.06	29.54
Web-ID	**51.88**	**48.32**	**37.49**	29.49

Table 2: WERs according to the number of corrective actions, for initial transcriptions of WERs above 40%.

than a simple left-right method (human only - without interactive decoding). The semantics-based strategies yield better results, especially the Web-based approach (Web-ID). Nevertheless, they are slightly less efficient than the interactive left-right method (LR-ID), which is probably comfortable for the transcriber.

Table 2 presents results obtained with the massively erroneous class (WER $>$ 40%). The results are significantly different, semantic methods are, in this case, better than the other methods for the first ten correction actions. Especially, the Web-based approach which yields a strong gain on the first correction (-4.03% absolute). This method is the most efficient when it is applied to ten corrective actions, the gain is (-10.32% absolute) compared with the manual method.

Globally, the Web-based approach yields slightly better results than the corpus-based one, despite the fact that the test set (broadcast news) is closely related to the text corpus that we used (newswires). The benefits of the Web-based approach could be more important on tasks that are weekly covered by the used corpus.

5. Conclusion and perspectives

We presented and evaluated an interactive approach for the speech decoding which aims to minimise the overall cost of a transcription. The main idea is to alternate correction steps and decoding steps which that takes into account the transcriber corrections. Considering the correction in a particular area of a segment can become profitable on a reactive decoding, we proposed various strategies in order to drive the transcriptionist to the critical areas of the transcription. The results demonstrate that the interactive decoding provides a significant improvement of the correction efficiency, compared to the standard human-only technique. The comparison between the various transcriber driving strategies is more enlightening. Concentrating the corrective actions on the widest densities of the graph is by far less efficient than the simplest left-right strategy. One of the reasons is that a system in failure mode can exploit a low number of corrections to set the decoder on the straight and narrow path to success. Integrating a word in an area where the system hesitates cannot guarantee the efficiency of a block of words as well as the left-rigth correction. Semantic-based strategies do not yield a clear WER improvement for the lowest-WER segments, but are much more effective for the highest-WER class. These strategies allow to detect areas where corrections are the most efficient.

Moreover, semantic-based strategies could improve the semantic quality of the transcript, while reducing the number of required corrective acts. This could be an efficient way of correcting transcripts that are dedicated to speech indexing or understanding systems. We are now evaluating these approaches in this respect. Moreover, semantic-based strategies could improve the semantic quality of the transcript, while reducing the number of required corrective acts.

6. References

Thierry Bazillon, Yannick Estève, and Daniel Luzzati. 2008. Manual vs assisted transcription of prepared and spontaneous speech. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

G. Riccardi D. Falavigna, R. Gretter. 2002. Acoustic and word lattice based algorithms for confidence scores. pages 1621–1624.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. 2005. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In *European Conference on Speech Communication and Technology, Interspeech*, Lisbon, Portugal.

J.Ogata and M.Goto. 2005. Speech repair: Quick error correction just by using selection operation for speech input interfaces. In *International Conference on Speech Communication and Technology, Interspeech*, pages 133–136, Lisboa, Portugal.

Thomas Kemp and Thomas Schaaf. 1997. Estimating confidence using word lattices. In *Proc. Eurospeech '97*, pages 827–830, Rhodes, Greece.

Benjamin Lecouteux, Georges Linarès, J.F. Bonastre, and Pascal Nocera. 2006. Imperfect transcript driven speech recognition. In *Interspeech'06-ICSLP*, Pittburgh, Pensylvania, USA.

Benjamin Lecouteux, Georges Linarès, Yannick Estève, and Guillaume Gravier. 2008. Generalized driven decoding for speech recognition system combination. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Las Vegas, USA.

Georges Linarès, Dominique Massonié, Pascal Nocera, and Christophe Lévy. 2007. The lia speech recognition system : from 10xrt to 1xrt.

Hiroaki Nanjo, Yuya Akita, and Tatsuya Kawahara. 2006. Computer assisted speech transcription system for efficient speech archive. In *Western Pacific Acoustic conference*, Seoul, Korea.

C. Van Rijsbergen. 1979. Information retrieval. MA, USA.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, Colorado, USA.