Assessing the effectiveness of slides as a mean to improve the automatic transcription of oral presentations

A.Peregoudov [a b] A.Vinciarelli [a b]
H.Bourlard [a b]

IDIAP–RR 06-56

October 2006

submitted for publication

[a] IDIAP - {peregoud,vincia,bourlard}@idiap.ch
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (Switzerland)

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-56

Assessing the effectiveness of slides as a mean to improve the automatic transcription of oral presentations

A.Peregoudov A.Vinciarelli H.Bourlard

October 2006

submitted for publication

Abstract. This paper presents experiments aiming at improving the automatic transcription of oral presentations through the inclusion of the slides in the recognition process. The experiments are performed over a data set of around three hours of material (∼33 kwords and 270 slides) and are based on an approach trying to maximize the similarity between the recognizer output and the content of the slides. The results show that the upper bound to the Word Error Rate (WER) reduction is 1.7% (obtained by transcribing correctly all words co-occurring in both slides and speech), but that our approach does not produce statistically significant improvements. Results analysis seems to suggest that such results do not depend on the similarity maximization approach, but on the statistical characteristics of the language.

1 Introduction

The automatic transcription of oral presentations is a cornerstone in any approach trying to capture, store and index the information delivered by the speakers. On the other hand, the recognition of spontaneous speech is still a challenge and the average Word Error Rate (around 30%) is still higher than in other kinds of data such as recordings of news professional speakers (around 5-10%). The goal of this work is to reduce the WER of presentation transcritpions by maximizing the similarity between the output of the recognizer and the text of the slides. The rationale behind such an approach is that slides and speech carry the same message, then they are supposed to be similar.

Our experiments involve three hours of talks for a total of around 33,000 words and 270 slides. We perform two sets of experiments using the same similarity maximization technique (see Section 3): the first experiments set assumes unrealistically that the words co-occurring in both speech and slides are known in advance (this enables us to estimate the highest possible improvement with our approach). The second experiments set uses the same technique in realistic conditions (words co-occurring in speech and slides are not known in advance) and it measures the actual decrease of the WER that can be achieved over our data.

The results show that the WER cannot be decreased by more than 1.7% (when knowing in advance the words co-occurring in speech and slides) and that the application of the similarity maximization in realistic conditions does not produce statistically significant WER changes. The analysis of the results seems to suggest that this does not depend on the techniques we propose, but rather on the statistical characteristics of language. We propose in particular two possible explanations: the first is the language *sparseness*, i.e. the fact that most words occur few times (less than 4) and account for a small fraction of the total number of words appearing in a text (see Section 4). The second is the statistical independence between the event of a word being uttered and the event of a slide being projected (see Section 4). Both above effects account for the small number of words co-occurring in slides and speech, then for the low WER impact of similarity maximization.

The rest of this paper is organized as follows: Section 2 presents the experimental setup, Section 3 presents our approach, Section 4 presents experiments and results and Section 5 draws some conclusions.

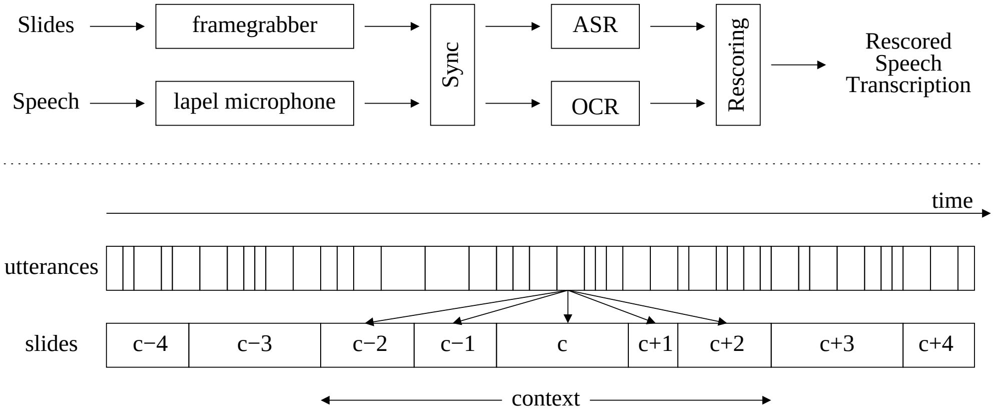

Figure 1: Experimental setup. The scheme shows the experimental setup of our system: slides and speech are captured using different devices and synchronized. They are then submitted to the recognition systems and given as input to the rescoring before giving the recognition output.

2 Data Collection

The data is captured online during the presentations, as illustrated in the upper part of Figure 1. The speakers are equipped with lapel microphones which acquire the speech signal. The slides projected on the screen through the PC-projector are captured with a framegrabber, i.e. a hardware device which converts the analog video stream of the projector into a sequence of digital images. The capture is performed once per second and the resulting image sequence is processed to detect slide transitions (see [4] for the advantages of such an approach).

A state-of-the-art Automatic Speech Recognition (ASR) system (with a dictionary of 50,000 words) [2] converts the speech signal into a *confusion network* (see Section 3), i.e. a graph of word hypotheses that can be used to extract the word sequence best matching the acoustic observations. At the same time, an advanced Optical Character Recognition (OCR) system [4], robust to complex background and font variability, detects and recognizes the text in the slide images. The synchronization enables one to know the slide displayed at the moment a given sentence is uttered. The result is a sequence of I utterances $u^{(i)}$ synchronized with a sequence of J slide segments $s^{(j)}$ (the slide segment is the time interval during which a certain slide is projected). The slide $s^{(c)}$ displayed during every utterance $u^{(i)}$ defines a *slide context* of radius R, i.e. a set $C = \{s^{(c-R)}, \dots, s^{(c+R)}\}$ (see lower part of Figure 1) of slides centered around the utterance $u^{(i)}$.

The experiments are performed over 6 presentation recordings from the AMI data corpus (`mmm.idiap.ch`). The presentations have a total duration of approximately 3 hours. The total number of words in speech and slides is 32,960 and 7,515 respectively. The WER of ASR and OCR are 32.5% and 20.0% respectively.

3 Similarity Maximization

This section shows how we maximize the similarity between the recognizer output and the texts of the slides belonging to the context C of the utterance being recognized. The approach we propose is based on the so-called *confusion networks* [3], i.e. graphs where each node corresponds to a moment in time and each edge corresponds to a transcription hypothesis for the time interval between the two connected nodes (see Figure 2). The confusion networks are extracted from the lattices output by the recognizers and provide a probability for each word hypothesis (see [3] for the details).

Given a time slot, $H_M = \{h_1, \dots, h_M\}$ is the set of the M most probable word hypotheses. If C is the context of the same time slot (see Section 2), then $H_S = \{w_1, \dots, w_N\}$ is the set of all words appearing in the C slides. In order to maximize the similarity between the slides and the recognizer output, we select as a transcription of the time slot the word with the highest probability in the following set H:

$$H = \begin{cases} H_M & \text{if} \quad H_M \cap H_S \quad \text{is empty} \\ H_M \cap H_S & \text{otherwise,} \end{cases} \tag{1}$$

in other words, whenever possible, the list of the hypotheses is limited to the words co-occurring in both slides and speech.

The above approach is not applied to any time slot. In fact, for some slots H_M contains only *stopwords*, i.e. words belonging to a list (called *stoplist*) which contains articles, pronouns, prepositions and other content neutral words necessary to build gramatically correct sentences. If this is the case, then the above similarity maximization approach is not applied. In our experiments, the stoplist contains 389 entries. The words that are not in the stoplist are called *content words*.

4 Experiments and Results

This section presents the experiments performed in this work and two possible explanations for the results we obtain. We perform two sets of experiments: the first set involves an *Oracle*, i.e. a system which knows in advance when the words written in the slides are uttered. Such a system is not

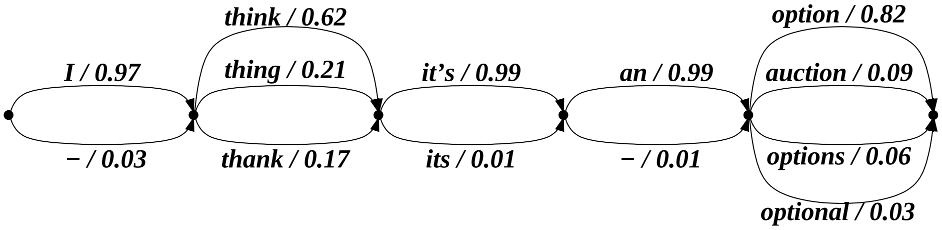

Figure 2: Example of a confusion network. Edges are labelled with hypotheses and corresponding posterior probabilities."-" stands for a deletion. The details on the estimation of the hypothesis probabilities are given in [3].

realistic, but it provides the largest possible improvement we can obtain by applying the similarity maximization technique described in Section 3. The Oracle is implemented by forcing the confusion networks to allocate a slot for each H_S content word uttered by the speakers and by using the correct transcription as the only hypothesis for such slots. The second set of experiments applies the technique of Section 3 in a realistic setting, i.e. without knowing in advance which words co-occur in slide and speech. These experiments provide the actual results we achieve over our data.

The results are shown in Figure 3. The dotted line corresponds to the performance of the recognizer without the application of the similarity maximization approach. The Oracle results show that, by increasing the context radius, the WER is decreased by up to 1.7% (statistically significant) by including all the slides of a given presentation in the context of each utterance. The situation is different for the realistic system. Figure 3 reports the results when keeping only two hypotheses per slot (see Section 3) and when keeping all the hypotheses per slot. The increase of the context radius increases the WER in both cases, but in the first case (two hypotheses) the degradation becomes statistically significant only when all slides are included in the context.

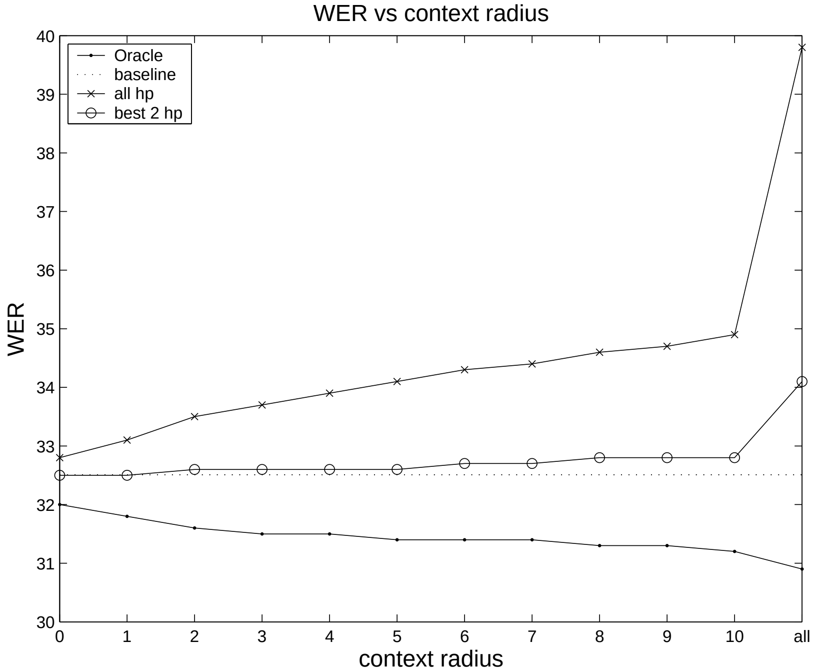

Figure 3: Results. The plots show the WER as a function of the context radius R.

The above results seem to contradict the assumption that speech and slides carry the same message, but we propose two possible explanations for the apparent mismatch between speech and slides. The first is based on the *Zipf's Law* and shows that most content words occur too rarely (less than 4 times) to actually influence the WER. The second is based on a statistical test which shows how the event of uttering a word and the event of a certain slide being projected are statistically independent. This means that the slides, although carrying the same content of the speech, do not affect the distribution of the uttered words to an extent sufficient to impact the WER. The next two subsections present in more details the above explanations.

4.1 Sparseness and Zipf's Law

The main effect of the language sparseness can be observed in Figure 4. The plot shows the so-called *Zipf's Law* [5]:

$$n_r \sim \frac{1}{rank(r)} \tag{2}$$

where n_r is the number of words appearing r times and the function at the denominator is the rank of r among the observed number of occurrences. The data in the plot are extracted from the Wall Street Journal (WSJ) Corpus, one of the main benchmark of the Information Retrieval literature, but the conclusions apply to any collection of texts. The Zipf's Law shows that around half of the unique words appearing in a text corpus occur less than 4 times and represent around 1% of the word mass (the total number of words in the corpus). If the words occurring up to twenty times are considered, we obtain more than 85% of the lexicon, but just 4% of the word mass. These are the words that actually characterize the content of a text and they represent only a minor fraction of the total word mass. This means that even if slides and speech carry the same message, the fraction of shared content words, i.e. the fraction of words that can be corrected through a rescoring approach, is small. The corresponding effect on the WER is then small as well.

4.2 Slides-Speech Statistical Independence

The second possible explanation for the rescoring results can be obtained by testing the statistical independence between the event of a certain slide being projected and the event of a certain word being uttered. If $p(w_i, s_j)$ is the probability of uttering word w_i when the slide s_j is projected, then the above corresponds to test the following hypothesis:

$$\mathbf{H_0}: \quad p(w_i, s_j) = p(w_i)p(s_j), \tag{3}$$

where $p(w_i)$ is the probability of word w_i being uttered and $p(s_j)$ is the probability of slide s_j being projected. If a_{ij} is the number of times word w_i is uttered when slide s_j is being displayed, then the Maximum Likelihood estimation of the above probabilities is as follows:

$$p(w_i, s_j) = \frac{a_{ij}}{\sum_{ij} a_{ij}}; \quad p(w_i) = \frac{\sum_j a_{ij}}{\sum_{ij} a_{ij}}; \quad p(s_j) = \frac{\sum_i a_{ij}}{\sum_{ij} a_{ij}}. \tag{4}$$

Since most of the a_{ij} counts are null, a smoothing approach is applied in order to redistribute the probability mass over unseen events [1].

We tested the independence (using a Pearson's χ^2 test [1]) hypothesis for all presentations at disposition and for both manual and automatic transcriptions of slides and speech. The result is always that the hypothesis is true with a confidence level of 99%. In other words, the fraction of words for which the distribution is influenced by the slide being projected is negligible. This means that the fraction of words that can be corrected using the approach of Section 3 is too small to influence significantly the WER.

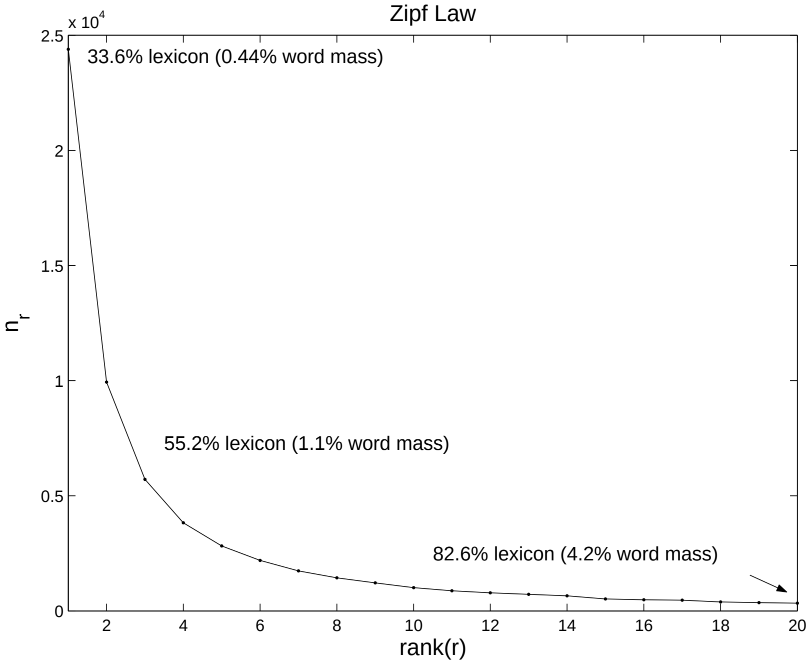

Figure 4: Zipf's Law. The plot shows the number of words appering r times as a function of r (since all r values are represented, the value of r corresponds to the rank of r among the represented number of occurrences).

5 Conclusions

This paper has presented experiments aimed at improving the automatic transcription of slide based oral presentations by including the slides in the recognition process. The results obtained with an Oracle system, i.e. with a system which *knows* when a word written on the slides is uttered by the speaker, show that the highest possible improvement of the WER is 1.7%. On the other hand, the results obtained in realistic conditions show a degradation (although non statistically significant) of the WER. We propose two possible explanations suggesting that the above results do not depend on the similarity maximization approach, but rather on the statistical characteristics of language: the first is the language sparseness, i.e. the fact that most words, especially those that are more content specific, account for a small fraction of the word mass (around 4%). The second is the statistical independence between the event that a certain slide is being projected and the event that a certain word is uttered. Both above explanations seem to suggest that the problem is not in the approach we use to perform the similarity maximization, but in the in the fact that two texts, although being related, share too few words to provide helpful information about each other.

On the other hand, there can be some exceptions to such a conclusion, e.g. some speakers read the content of the slides when presenting and this can enhance the recognition results if the slides text is included in the corpus used for training the Language Models. In other cases, the slides can contain many proper names or acronyms that are unlikely to be in the dictionary of a recognizer and when they occur enough, the impact on the WER can be more important. However, the above situations seem to be the exception rather than the rule.

References

[1] R. Christensen. *Log-linear models and logistic regression*. Springer Verlag, 1997.

[2] T. Hain, L. Burget, J. Dines, I. McCowan, M. Karafiat, M. Lincoln, D. Moore, G. Garau, V. Wan, R. Ordelman, and S. Renals. The development of the AMI system for the transcription of speech in meetings. In *Proceedings of the conference on Machine Learning in Multimodal Interfaces*, 2005.

[3] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400, 2000.

[4] A. Vinciarelli and M. Odobez, J. Application of information retrieval technologies to presentation slides. *IEEE Transactions on Multimedia*, 8(5):981–995, 2006.

[5] G.K. Zipf. *Human behavior and the principle of least effort*. Addison Wesley, 1949.