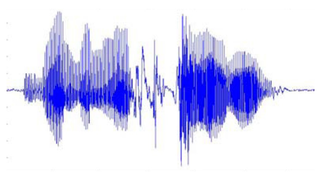
Imperfect transcript driven speech recognition

Benjamin Lecouteux, Georges Linares, Pascal Nocera, Jean-François Bonastre

Imperfect transcript driven speech recognition

Benjamin Lecouteux, Georges Linarès, Pascal Nocéra, Jean-François Bonastre
Université d'Avignon - {benjamin.lecouteux, georges.linares, pascal.nocera, jean-francois.bonastre}@univ-avignon.fr

- actor
- journalist



audio stream

Speech recognizer (based on A* algorithm)

Automatic transcript aligned and corrected

- prompter texts
- scenario scripts

On-demand synchronization of the A* search with imperfect transcript by the DTW algorithm

Generic language model (65K words)

LM trained on imperfect transcript

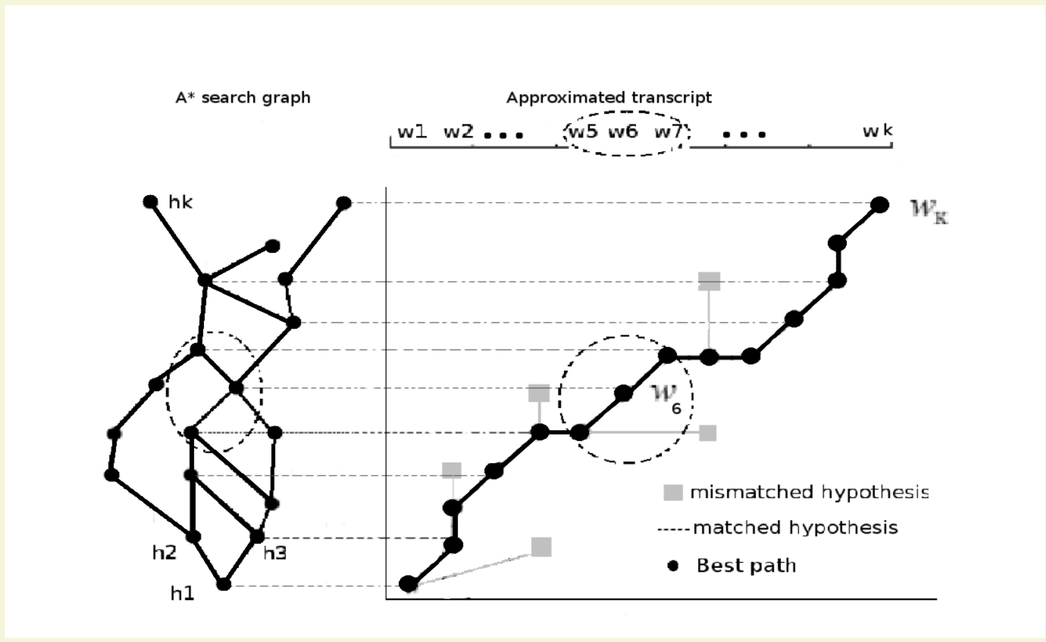
Approximated transcript (~10% WER)

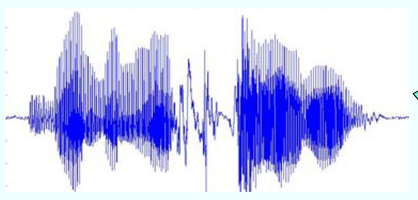
70% 30%

Interpolated language model

$$P'(w|h) = \sum_{i=1}^{k} \alpha_i P_i(w|h)$$

On-demand synchronization and estimate function rescoring





Audio stream

Automated transcript aligned and corrected

Alignment with A algorithm during asynchronous decoding*

- Each evaluated word is aligned to the reference word stream using a Dynamic Time Warping (DTW) algorithm.

- Once the hypothesis is synchronized with the transcript, the algorithm estimates the matching transcript-to-hypothesis score α .

- Then, linguistic probabilities $P(w_i|w_{i-2}, w_{i-3})$ are modified using the following rescoring rule:

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

- α is maximum when the trigram is aligned and decreases according to the misalignments of the history.

Alignment example.

Prompter : france inter flash d'informations à huit heures un quart
Pronounced text : france inter l' actualité à huit heures un quart

Results after decoding :

Without alignment : FACE À ELLE actualité à huit heures ET quart

With alignment : france inter l' actualité à huit heures un quart

Experimental context :

- Experiments assessed on 3 hours of radio broadcast

- 10 % WER introduced in transcripts

- Generic language model used : 65000 words trained on "Le Monde"

- Speech recognition system : SPEERAL, an asynchronous decoder based on the A* algorithm.

Results :

- Initial decoding : 22,7% WER using a generic language model.

ML-G : Generic model language 65K words
ML-TrErr : Language model trained on the transcript
alTrEr : Alignment to the imperfect transcript

- Decoding using LM interpolation and alignment.

	$\overline{WER}$
ML-TrErr + *alignment* TrEr	9.9%
ML-G + *alignment* TrEr	7.7%
ML-G70%+ML-TrEr30%+alTrEr	7.2%
ML-G50%+ML-TrEr50%+alTrEr	7.4%
ML-G30%+ML-TrEr70%+alTrEr	8.6%

- Decoding using language model interpolation.

	Taux d'erreur
ML-TrErr seul	16.3%
ML-G 70% + ML-TrErr 30%	16.2%
ML-G 50% + ML-TrErr 50%	15.4%
ML-G 30% + ML-TrErr 70%	15.2%

Conclusions :

- Best results : interpolation 70-30% with forced alignment to transcript
- WER better than the transcript : 7.2%
- **relative improvement : 28%**