An Automatic Lightly Supervised Speech Segmentation and Alignment Tool

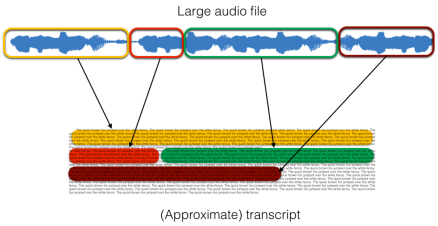
A. Stan[a], Y. Mamiya[b], J. Yamagishi[b,c], P. Bell[b], O. Watts[b], R.A.J. Clark[b], S. King[b]

a) Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., Cluj-Napoca, 400027, Romania
b) The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
c) National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Computer Speech and Language, vol. 35, pp. 116-133, 2016 Download link: http://simple4all.org/product/alisa/

Objective: Aligning large audio files with their approximate transcripts at sentence-like chunk level

Why? To obtain training material for speech enabled applications without the need to record or transcribe the data.



Existing solutions

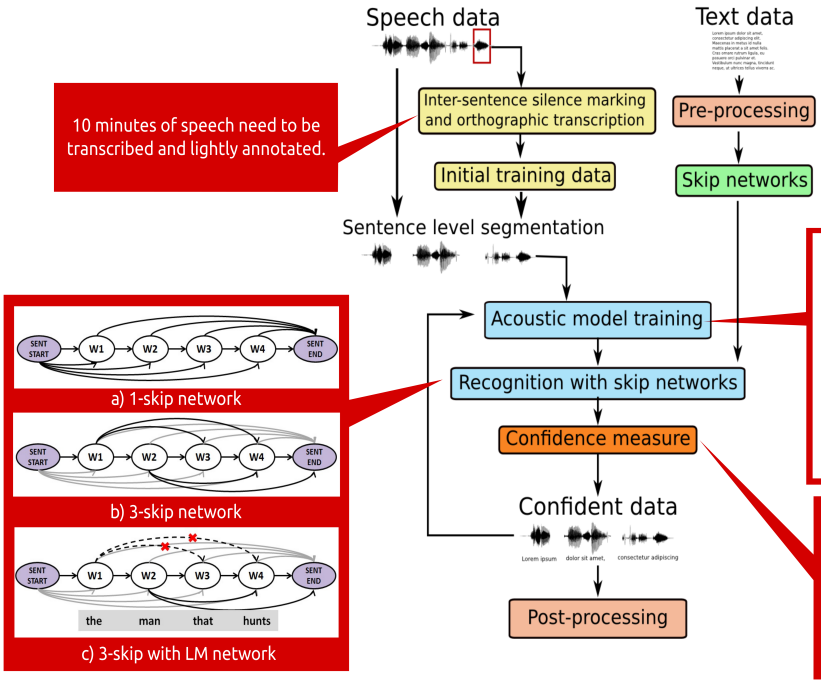
When the text matches the speech data entirely:

* forced alignment;
* dynamic time warping;
* modified Viterbi algorithms.

When there are imperfections in the transcript:

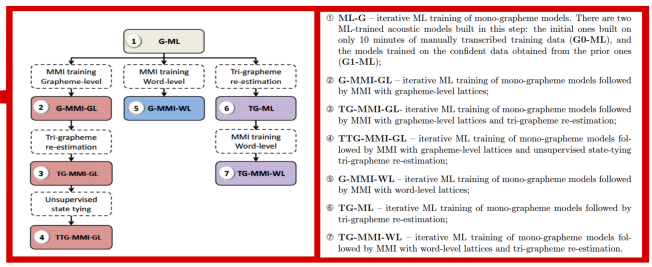
* factor automaton;
* good acoustic and linguistic models;
* phone-level acoustic decoders.

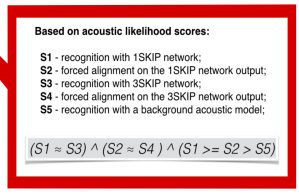
HOW?



Requirements

* Any (alphabetic) language;
* Minimum user intervention;
* No prior expert knowledge is available;
* The resource does not have to be transcribed entirely;
* The text transcript is chronologically correct.





Plenty of speech resources and their (*approximate*) transcripts are available on the internet

Results

Acoustic model	All data SER [%]	All data WER [%]	Confident data Percent [%]	Confident data SER [%]	Confident data WER [%]
① G0-ML	50.56	5.50	48.23	12.14	0.74
① G1-ML	47.20	5.12	56.98	11.15	0.58
② G-MMI-GL	56.30	5.94	71.63	18.74	1.42
③ TG-MMI-GL	24.41	2.67	78.55	7.42	0.48
④ TTG-MMI-GL	21.82	2.30	79.34	6.83	0.44

Table 1: Error rates for the objective evaluation of the alignment method for the English audiobook. The All Data results are reported using the 3SKIP with LM network.

Acoustic model	All data SER [%]	All data WER [%]	Confident data Percent [%]	Confident data SER [%]	Confident data WER [%]
① G0-ML	79.65	19.29	44.82	27.50	7.75
① G1-ML	52.79	6.98	51.66	12.76	0.62
② G-MMI-GL	54.32	6.17	53.50	19.50	0.96
③ TG-MMI-GL	49.33	10.84	57.51	14.22	0.69
④ TTG-MMI-GL	27.46	3.38	65.69	7.22	0.30

Table 2: Error rates for the objective evaluation of the alignment method for the French audiobook. The All Data results are reported using the 3SKIP with LM network

System	SER [%]	WER [%]
ASR unadapted, general LM	78.54	13.89
ASR adapted, general LM	74.25	11.29
ASR adapted, biased LM	24.48	2.18
ALISA	21.82	2.30
ALISA supervised lexicon	18.57	2.12

Table 3: Error rates of the entire speech data for state-of-the-art ASR system unadapted witha general language mode, state-of-the-art ASR system with adapted acoustic models, state-of-the-art ASR system with adapted acoustic models and biased language model, ALISA and ALISA with a supervised phonetic lexicon. The ALISA results are reported using the 3SKIPwith LM network.

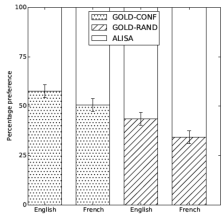


Fig. 1. AB preference score results of the listening tests for English and French. ALISA systems use the training data obtained with fully automatic segmentation and alignment. GOLD-CONF systems use the same utterances as ALISA but use a gold-standard segmentation and alignment. GOLD-RAND systems use a random selection of utterances from the gold-standard segmentation and alignment of the same duration as the ALISA systems.

The GOLD transcripts for A Tramp Abroad were kindly provided by Toshiba Research Europe Limited, Cambridge Research Laboratory.
The GOLD transcripts for Candide ou l'optimisme were provided by CCProse.
The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 287678 (Simple4All), as well as by EPSRC EP/I031022/1 (NST) and EP/J002526/1 (CAF).
The work presented here has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF: http://www.ecdf.ed.ac.uk). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk).