Distant Speech Recognition in a Smart HomeComparison of Several Multisource ASRs in Realistic Conditions

Benjamin Lecouteux, Michel Vacher, François Portet

Distant Speech Recognition in a Smart HomeComparison of Several Multisource ASRs in Realistic Conditions

Benjamin Lecouteux, Michel Vacher, François Portet

Sweet-home project and corpora

Audio-based interaction technology :
- to provide assistance by natural man machine interaction
- to ease social inclusion
- to provide security reassurance by detecting situations of distress

- 2 phases and 21 speakers

- Phase 1: participants uttered 40 predefined casual sentences on the phone but were also free to utter any sentence they wanted.

- Phase 2 consists in reading aloud a list of 44 sentences whose 9 were distress sentences and 3 were domotic orders.

- Phase 1 : 862 sentences (38 minutes 46s per channel in total) → used for supervised MLLR speaker adaptation and dev

- Phase 2 : 917 sentences (40 minutes 27s per channel in total).
for Phase 2 all from 21 speakers → used for test

Approaches

Beamforming

- Based on the weighted&sum microphone array theory.

- The 7signals are entirely combined for each speaker.

- Involves low computational cost and combines efficiently acoustic streams to build an enhanced acoustic signal.

ROVER

- Baseline ROVER was tested using all available channels without a priori knowledge.

- In a second time, an a priori confidence measure based on the SNR.

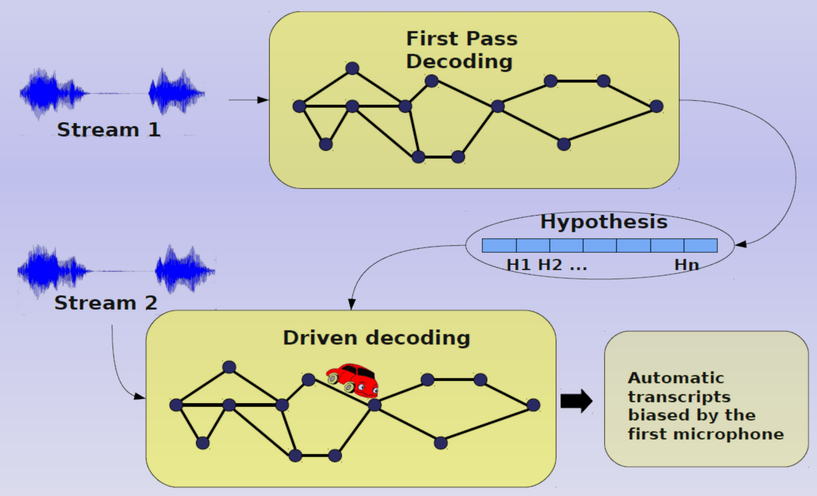
Driven Decoding Algorithm (DDA)

- DDA aims to align and correct auxiliary transcripts by using a speech recognition engine.

- We propose to use a variant of the Driven Decoding Algorithm where the output of the first microphone is used to drive the output of the second one.

- The second ASR system speed is boosted by the approximate transcript (only 0.1xRT).

- DDA merges truly and easily the information from the two streams while voting strategies (such as ROVER) do not merge ASR systems outputs.



Home automation and extended DDA

Detection of home automation and distress sentences

- Each order and distress sentences are transcribed in phoneme graph in which each path is a variant of pronunciation.

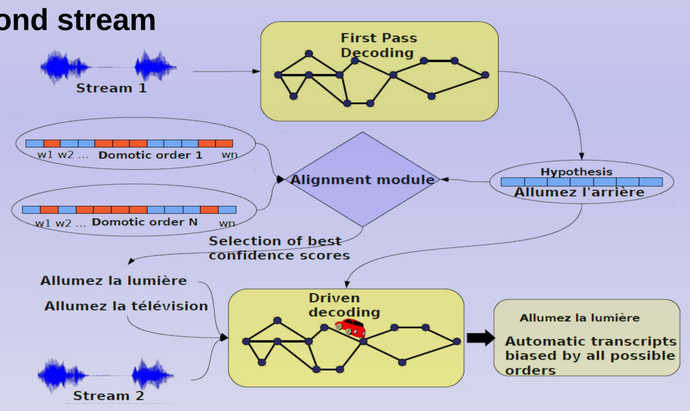
- In order to locate domotic orders into automatic transcripts each sentence from domotic orders are aligned by using a Dynamic Time Warping.

-Each domotic order is aligned and associated with an alignment score: the percentage of well aligned symbols.

-This approach takes into account some recognition errors such as word endings or slight variations.

DDA extension by using predefined sentences

- DDA 2-level: vocal orders are recognized from the first decoded stream which are then used to drive the decoding of the second stream



- Speech segments of the first pass are projected into the 3 − best vocal orders by using an edit distance and injected via DDA into the ASR system for the fast second pass.

Results

Wer, domotic orders detection

Method	WER$^{\pm SD}$	Domotic recall	Domotic precision	F -measure
Baseline	$18.3^{\pm 12.1}$	88.0	90.5	89.2
Oracle Baseline	$17.7^{\pm 10.3}$	88.5	91.3	89.9
Beam Forming	$16.8^{\pm 8.3}$	89.0	92.6	90.8
DDA +SNR	$11.4^{\pm 5.6}$	93.3	97.3	95.3
DDA 2 lev.+SNR	$\mathbf{8.8}^{\pm 3.7}$	**95.6**	**98.1**	**96.8**
ROVER	$20.6^{\pm 8.5}$	85.0	90.0	87.4
ROVER 2c+SNR	$13.0^{\pm 6.6}$	91.3	95.3	93.3
ROVER +SNR	$\mathbf{12.2}^{\pm 6.1}$	**92.7**	**97.4**	**95.0**
ROVER Oracle	$7.8^{\pm 2.7}$	99.4	98.9	99.1

- Results show that combining all channels increases the ASR task robustness.

- DDA method showed a 37.8% relative improvement by using the SNR.

- The 2 level DDA presented a 52 % relative improvement.

- ROVER and the two DDA configurations led to a significant improvement.

Conclusion and perspectives

- Results confirmed that the use of the seven microphones improved the ASR accuracy.

- Beamforming improved the WER (16.8%), however its performance were very close to the baseline one (18.3%).

- The Driven Decoding Algorithm gave the best performance with a 11.4% WER and 95.3% F-measure for vocal order classification.

- The Driven Decoding Algorithm gave the best performance with a 11.4% WER and 95.3% F-measure for vocal order classification.

- This study shows that good recognition rate can be obtained by adapting classical ASR systems mixing multisource and domain knowledge.

- We plan to adapt these approaches to noisy conditions notably by applying source separation techniques to real daily living records composed of uncontrolled noise.