Selection and Combination of Hypotheses for Dialectal Speech Recognition

Victor Soto, Olivier Siohan, Mohamed Elfeky, Pedro Moreno

Columbia University & Google Inc., New York

Main Result

- Two methods to select and combine the best decoded hypothesis from a pool of dialectal recognizers are proposed.
- Machine Learning approach using features extracted from the ASR pipeline along with Word Embeddings.
- Experiments show very significant improvements for the selection scheme.

Dialectal Speech Recognition

Dialects are variations of the same language, specific to geographical regions or social groups. Differentiated at various linguistic levels:

- **Pronunciation**: water in SAE vs. British English
- **Orthographical**: color vs. colour
- **Vocabulary**: cell vs. mobile

Building a global ASR to decode dialectal variations has been shown to underperform. Building dialect-specific recognizers works best, but there is large variance in performance depending on size and quality of dialectal data, etc.

Question: How can we make use of a pool of dialectal speech recognizers to improve dialectal speech recognition?

1. **Cross-dialect** experiments show that on average best performance on a test set is always obtained by the dialectal-specific ASR.
2. **Hypothesis Selection Oracle** experiments show that there is room for large WER improvements if we learn how to choose which ASR to decode.
3. **Hypothesis Combination Oracle** experiments show that there is even more room for improvement if we use every dialectal ASR, combine their 1-best hypothesis and learn to choose word candidates.

	Production ASRs				Oracles	
Dataset	Egyptian	Gulf	Levantine	Maghrebi	Selection	ROVER
Egyptian (D)	**37.4**	43.5	44.3	53.1	26.9 (+28.1%)	23.1 (+38.2%)
Egyptian (VS)	**34.7**	38.2	42.2	48.2	23.6 (+47.0%)	19.4 (+44.1%)
Gulf (D)	36.2	**29.4**	34	47.4	20.8 (+29.3%)	18.7 (+36.4%)
Gulf (VS)	27.6	**21.5**	26.3	37.3	14.3 (+33.5%)	12.7 (+59.1%)
Levantine (D)	41.2	38	**33.7**	48.9	25.7 (+23.7%)	23.1 (+31.5%)
Levantine (VS)	34.7	29.9	**28.4**	41	19.9 (+29.9%)	17.7 (+37.7%)
Maghrebi (D)	44.2	41.5	41.6	**38.4**	24.6 (+35.9%)	21.1 (+45.1%)
Maghrebi (VS)	42.6	38.2	41.5	**34.7**	21.9 (+36.9%)	18.6 (+46.4%)

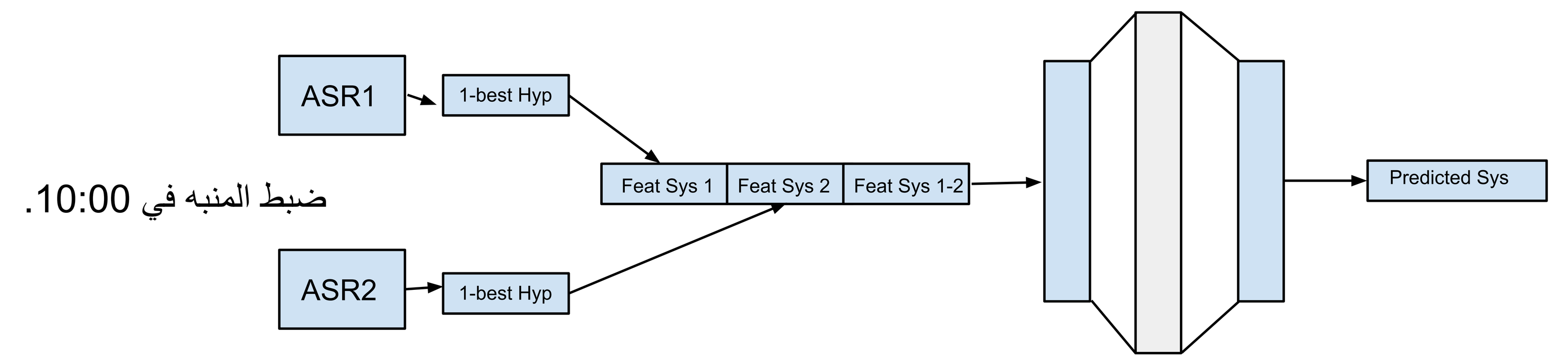
Left: cross-dialectal performance of each ASR (columns) in each dialectal test set (row). Right: oracle performance and relative improvements (Δ%).

Datasets

- Four dialect-specific corpora for **Egyptian, Gulf, Levantine and Maghrebi**.
- Train one ASR per dialect. 3M user utterances.
- DNN acoustic models (8 hidden layers, 1 bottleneck and 1 softmax layer). Input layer is 26 frames of 40-dim log-filterbanks each. Hidden layers have 2560 ReLU units each. Bottleneck has 256 linear activations and softmax layer holds 14336 units, one per CD state.
- ASR Test sets: one Google Voice Search (VS) and one Dictation (D) corpus per dialect. 25K utterances each.
- Hypotheses Selection and Combination experiments run using 5-fold cross-validation on test sets.

Hypothesis Selection

GOAL: To choose the hypothesis with the lowest WER.
HOW: Run all four dialectal ASRs (Egyptian, Levantine, Iraqi and Maghrebi) for each query, and use a ML classifier to predict best hypothesis.



FEATURE EXTRACTION

- Multi-label learning task (more than one ASR can have lowest WER).
- Utterance-level features: frame-averaged acoustic model cost, language model cost, minimum, maximum and average word confidence and word posterior; number of words, lattice density.
- Cross-system features: Levenshtein distance for each pair of hypotheses.
- Lexical features: later added bag-of-words embeddings (BWE) to our DNN input layer (64 dimensions).

CLASSIFIER

Feed-forward Neural Network with 1 hidden layer (512 ReLU units or 2048 when adding BWE) and an output layer of 4 Logistic Regression units.

Dataset	Best Hyp Selection	Δ%	+ BWE	Δ%
Egyptian (D)	36.1	+3.4	35.4	+5.3
Egyptian (VS)	31.8	+8.4	31.7	+8.6
Gulf (D)	28.6	+2.7	28.3	+3.7
Gulf (VS)	20.7	+3.7	20.4	+5.1
Levantine (D)	33.3	+1.2	33	+2.1
Levantine (VS)	26.4	+7.0	26.3	+7.4
Maghrebi (D)	34	+11.5	33.7	+12.2
Maghrebi (VS)	30.7	+11.5	30.5	+12.1

Left: WER performance using the baseline feature set. Right: WER results after adding the Bag-of-Words embedding (BWE) layer.

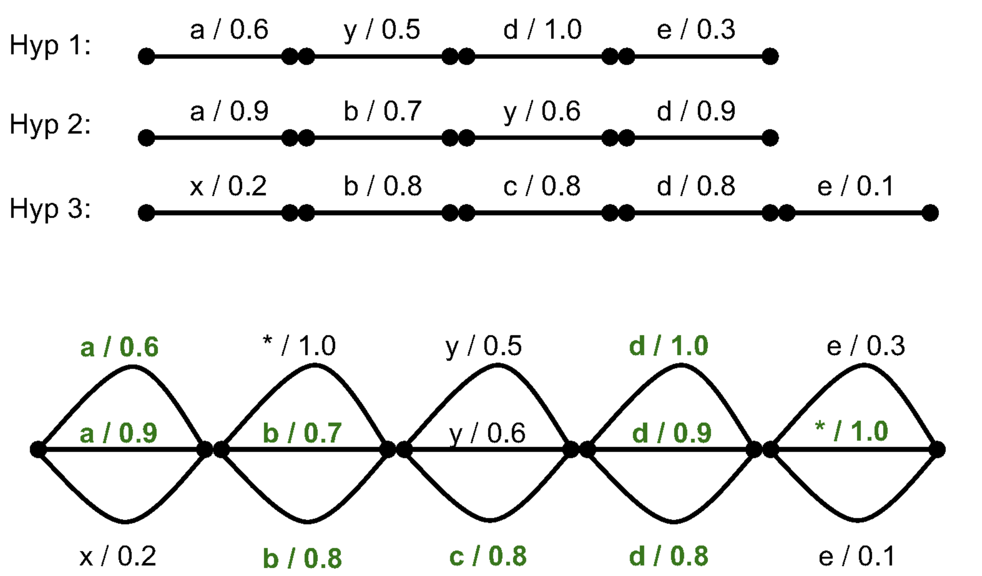
Hypothesis Combination

GOAL: Finding a word alignment of the dialectal hypotheses and selecting the correct arc (or epsilon) from each word bin.

WORD ALIGNMENTS

1. i-ROVER: ROVER Alignment + Best Arc Prediction using ML.
2. Label arc as correct or incorrect using true reference.

Example: for true reference ”a b c d”



FEATURE EXTRACTION

- Multi-label learning task (more than one word arc can be correct).
- Word-level features: acoustic model cost and language model cost of the FST arc and its frame-averaged values; weighted value of language and acoustic model costs; word confidence and lattice posterior; number of phones in the token; mean, std.dev., best, worst, and acoustic model scores at the frame level, and epsilon arc flag; lattice density within the token's time span.
- Lexical features: four layers of word embeddings, one per token.
- Contextual features: feature vectors of two previous word bins.

CLASSIFIER

Feed-forward Neural Network with 1 hidden layer (2048 ReLU units) and an output layer of 5 Logistic Regression units (one per token arc and epsilon/skip arc).

	ROVER		i-ROVER			
Dataset	Max.SumConf.	Δ%	iROVER	Δ%	+Context	Δ%
Egyptian (D)	38.4	-2.7	37.6	-0.5	37.6	0.0
Egyptian (VS)	34.5	+0.6	32.7	+5.8	32.9	-0.6
Gulf (D)	30.7	-4.4	29.8	-1.3	29.4	+1.3
Gulf (VS)	22.5	-4.7	21.2	+1.4	21	+0.9
Levantine (D)	34.6	-2.7	35.2	-4.5	35.2	0.0
Levantine (VS)	27.9	+1.8	27.6	+2.8	27.6	0.0
Maghrebi (D)	34.4	+10.4	35.3	+8.1	35.2	+0.3
Maghrebi (VS)	32.6	+6.1	31.2	+10.1	31.4	-0.6

ROVER (left subtable) and iROVER (right subtable) WER performance.

Conclusions

- Hypothesis selection scheme achieved between 1.2 and 12.1% relative WER improvements. Adding a word-of-bags embedding layer to the Neural Network further improved WER by 2.1 to 12.2%.
- Hypothesis combination (iROVER) with our own set of features and word embeddings. Got some improvements w.r.t baseline (1.4-10.1%) in some test sets, but underperformed in every test set when compared to the selection systems. Adding contextual features didn't help.