"There's no data like carefully selected data" — recognising disordered speech

Heidi Christensen, Iñigo Casanueva, Stuart Cunningham, Phil Green, Thomas Hain
University of Sheffield

Background

• People with dysarthric speech need models tuned to the specifics of their impairment, i.e., speaker-dependent (SD) models.

• There is rarely enough data to train SD models from scratch, but adapting from speaker-independent (SI) models have proven successful.

• Typically, *all* available speakers from relevant databases are included in this background SI model, but here we investigate whether this holds true for dysarthric speakers?

• For a given target speaker, should we be more discriminative and only include the most "similar" speakers? This would mean excluding speakers with impairments that are very different from those of the target speaker.

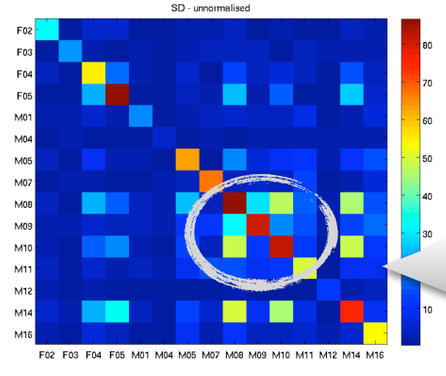
We test the hypothesis that selecting a subset of the available pool of speakers, based on how well they match the target speaker, can improve automatic speech recognition (ASR) accuracy.

• To test this we analyse how good a match each pool speaker is for a given target speaker (Q1),

• we propose and approach for selecting speakers based on this analysis and test whether we can improve on ASR accuracy this way (Q2)

• and finally, we explore different methods for automatically selecting the best pool (Q3)

Q1: Is there a large difference between how well the different speakers `match' each other?



Accuracies from testing each UAspeech speaker against the Speaker Dependent (SD) model of every other UAspeech speaker

There is clearly a difference between the different speakers

Q2: Can we improve on performance by only including some of the candidate speakers in the SI "pool" of speakers?

Approach

Algorithm 1 Training and testing SI($\mathcal{P}$)+MAP

1. Measure closeness of target speaker to all pool speakers
2. Rank pool speakers accordingly (rankedList)
3. Train all SI($\mathcal{P}$)+MAP models for different poolsizes, $\mathcal{P}$

for $\mathcal{P}$ =1 to $\mathcal{S}$ **do**
  curPoolList = { curPoolList, rankedList[$\mathcal{P}$] }
  Train SI(curPoolList)+MAP
  Accuracy [$\mathcal{P}$] = Test SI(curPoolList)+MAP
end for

Models and features:

• Features: 13-dim PLPs with first and second order derivatives.

• All HMMs trained using ML. State-clustered, triphones with Gaussian mixture models with 16 components pr state.

• MAP speaker adaptation.

Data:

• UAspeech database: Single-word recordings from 15 speakers with dysarthric speech, digits, Nato alphabet letters, common, command, and uncommon words.

Results:

Accuracies for varying poolsizes for speaker F04

The largest accuracy is obtained with only 4 speakers in the pool, i.e., excluding 10 data from 10 speakers from the model!

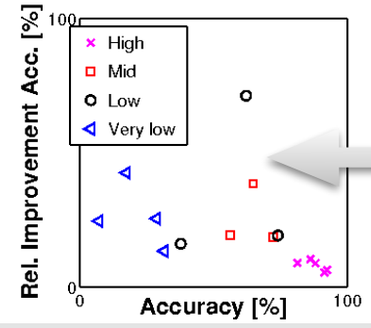
	Best Accuracy	Relative Improvement [%]	Best Poolsize	Amount of data
Avr.	59.1%	22	3.9	3.52

Q3: What's the best way of automatically ranking the candidate speakers?

Investigated different Ranking Methods: varying models and similarity measures

Method	-	-	RankMtd1	RankMtd2	RankMtd3	RankMtd4	RankMtd5	RankMtd6
Model/Measure	-	-	SD+MAP/Acc	SD+MAP/Acc	SD/lik'hood	SD/Acc	SI+MAP/Acc	tgtSpkSD/Acc
Final Model	SD	SI(all)+MAP	SI($\mathcal{P}$)+MAP	SI($\mathcal{P}$)+MAP	SI($\mathcal{P}$)+MAP	SI($\mathcal{P}$)+MAP	SI($\mathcal{P}$)+MAP	SI($\mathcal{P}$)+MAP
M04	6.6	6.5	8.1 (5)	7.8 (1)	6.8 (5)	7.3 (10)	**8.2 (1)**	6.0 (14)
F03	**32.1**	23.5	29.5 (4)	29.3 (2)	26.1 (7)	28.8 (4)	25.4 (2)	28.5 (3)
M12	15.1	11.5	16.4 (2)	14.6 (2)	12.8 (13)	17.6 (2)	**19.7 (2)**	14.3 (2)
M01	19.2	26.1	29.6 (3)	31.2 (3)	27.6 (6)	**31.6 (5)**	26.1 (12)	26.1 (14)
M07	62.1	61.3	73.1 (3)	72.6 (5)	70.2 (4)	**74.1 (3)**	70.7 (8)	70.5 (6)
F02	35.0	32.2	37.4 (6)	37.5 (3)	34.2 (7)	**37.8 (2)**	35.6 (2)	35.1 (11)
M16	56.4	35.3	60.3 (2)	57.9 (3)	53.8 (2)	62.2 (1)	**62.4 (1)**	47.7 (4)
M05	63.0	46.2	64.0 (2)	**65.3 (2)**	58.4 (3)	64.9 (3)	61.3 (1)	53.3 (3)
M11	54.4	49.3	58.8 (3)	**56.3 (3)**	52.6 (9)	**56.3 (3)**	53.2 (6)	51.2 (14)
F04	57.4	59.0	70.0 (4)	67.4 (6)	64.2 (2)	**72.4 (4)**	61.9 (3)	62.5 (12)
M09	82.4	78.2	86.3 (4)	**86.4 (3)**	85.6 (2)	86.3 (3)	83.4 (3)	83.9 (4)
M14	80.80	75.0	81.8 (4)	**81.6 (4)**	79.2 (2)	81.4 (3)	78.7 (2)	75.3 (6)
M10	83.5	85.8	90.5 (6)	89.0 (8)	88.1 (12)	**91.5 (5)**	90.0 (4)	87.3 (6)
M08	86.3	81.5	88.8 (3)	87.2 (6)	87.4 (1)	88.2 (2)	**88.4 (1)**	82.6 (10)
F05	89.2	87.0	92.6 (7)	**93.4 (7)**	91.0 (1)	92.5 (8)	91.3 (7)	90.4 (4)
Avr	54.9	50.6	59.1 (3.9)	58.5 (4.1)	55.9 (5.1)	**59.5 (4.1)**	57.1 (3.7)	54.1 (7.5)

RankMtd4 is best method with on average just over 4 speakers
(For comparison, CMLLR-SAT achieves 55.2 %)



`Mid', `low' and `very low' see biggest relative improvements. One `low' speaker had 60% relative improvement!

Conclusions

•Improvements in accuracies by on average 11.5 % compared to the SI, SD and CMLLR-SAT baselines.

•Speakers with moderate to severe impairments gain higher relative improvements.

•One speaker classed as having `low' intelligibility has over 60 % relative improvements

Dysarthric speech recognition is a domain with sparse resources; this research shows that despite this, we need to be careful when selecting data.

Acknowledgements:
This research was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology)