Modeling Accents for Automatic Speech Recognition

UNIVERSITY OF BIRMINGHAM

Maryam Najafian and Martin Russell, [mxn978, m.j.russell]@bham.ac.uk, School of Electronic, Electrical & Computer Engineering

1. Abstract

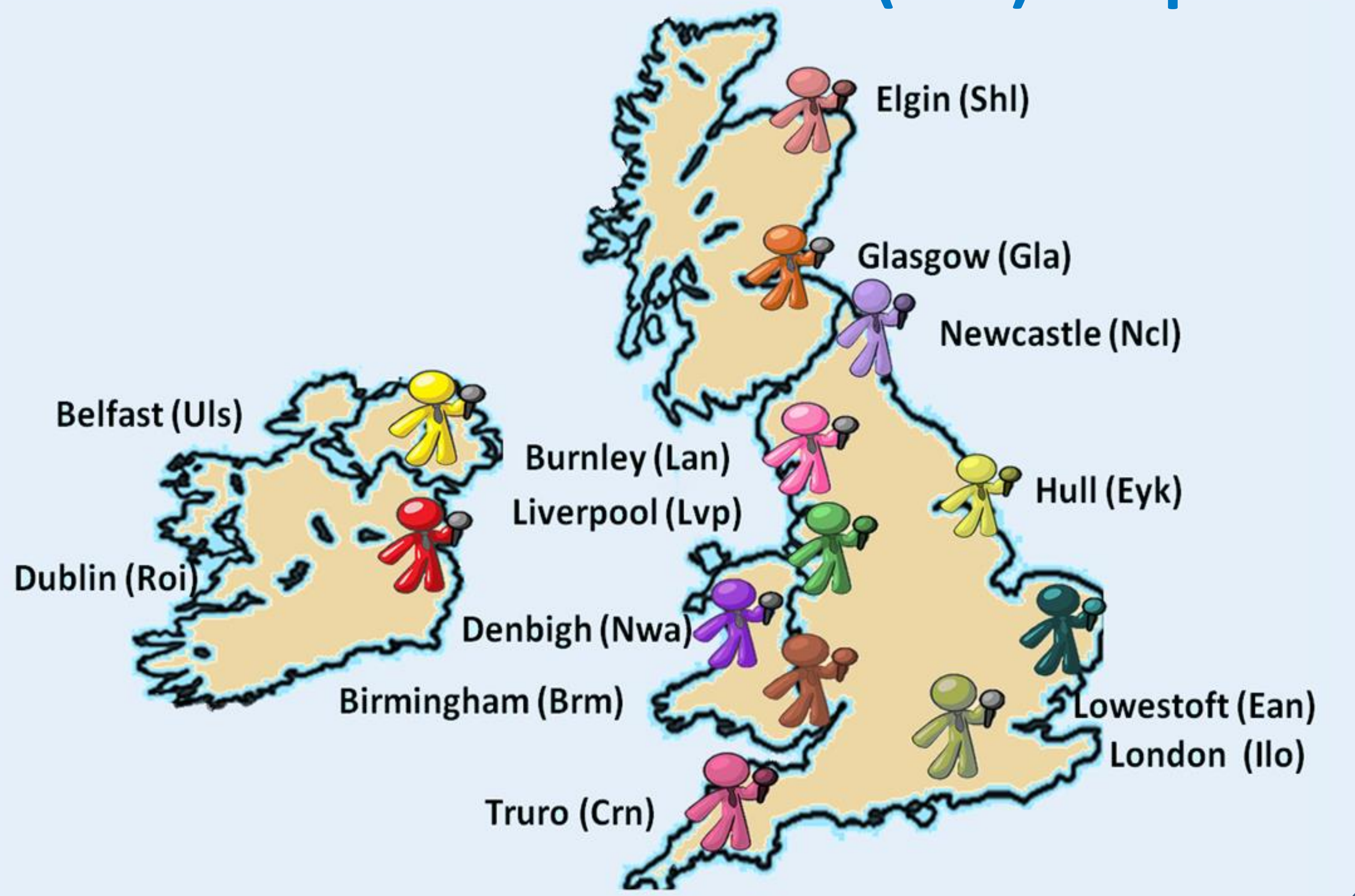
Automatic Speech Recognition (ASR) has many real-life applications.

Figure1. Current ASR Applications

Conventional adaptation techniques for ASR have two major limitations:

- They tend to ignore important factors including accents. Therefore, their performance is not consistent for speakers of different accents.
- They need a significant amount of training data from each individual to work well, but such data is not available in most real-life applications.
- This research is concerned with developing both rapid and robust ASR systems for British accents using two adaptation techniques namely, Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR) for adapting these systems to a new user using only 60 seconds of his/her speech.

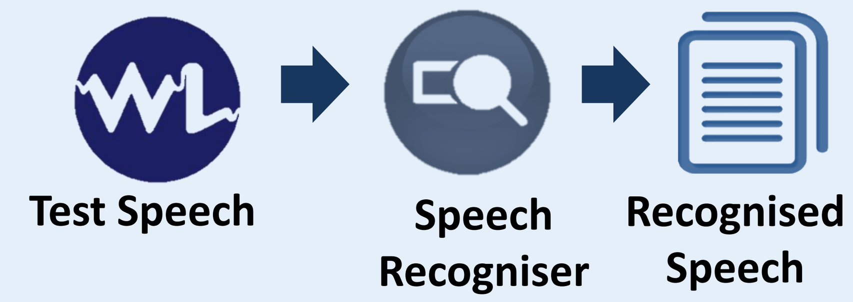
2. Accents of British Isles (ABI) Corpus



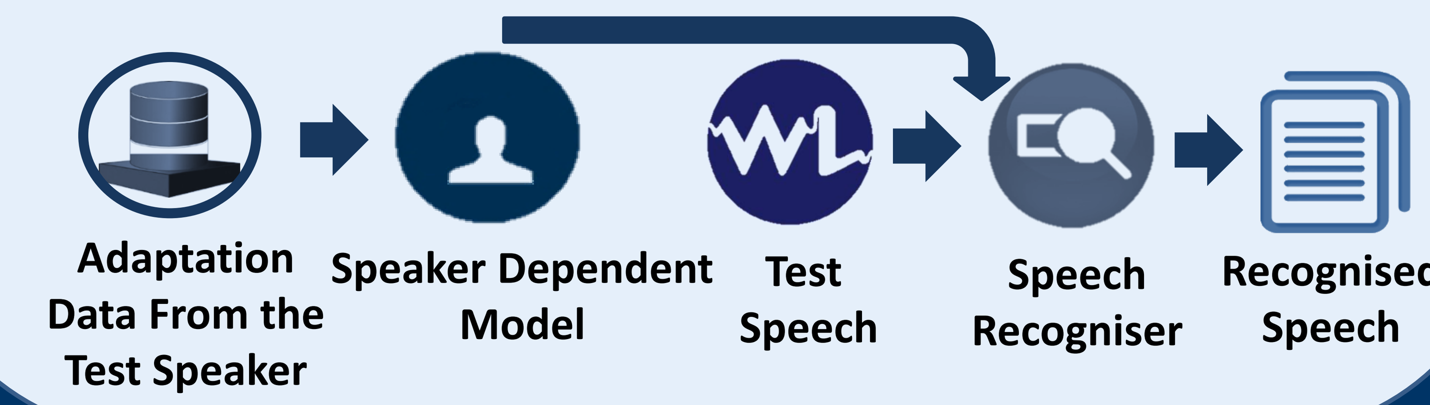
3. Methodology

- Methods EX0 and EX1 (below) show how current ASR systems work.
- Methods EX2 to EX4 show our proposed accent-dependent ASR model.
- In EX3 and EX4 Accent Distance Measure (ACCDIST) and in EX2 prior knowledge of test speakers accent is used For Accent Identification (AID) purpose.
- In EX5 all the models are adapted to the model from the SSE accent.

(EX0): Baseline experiment on the ABI corpus

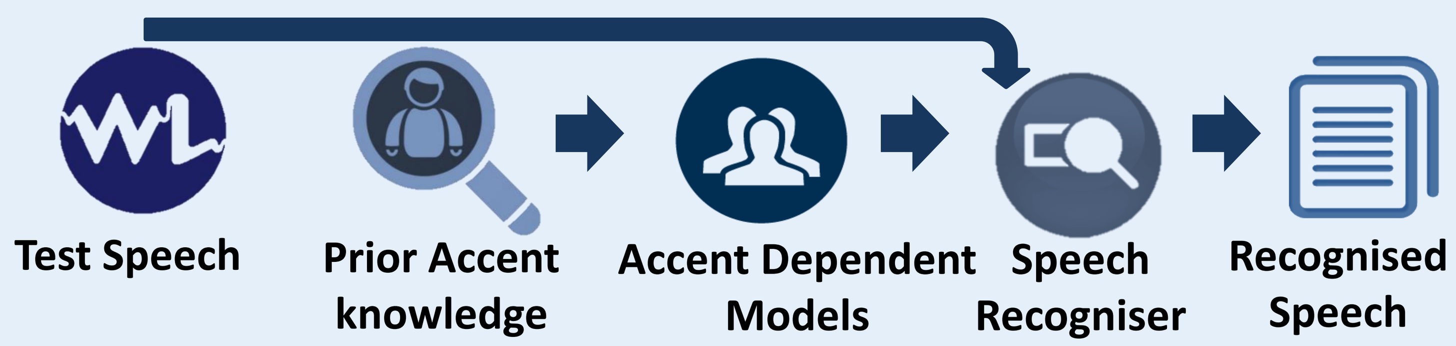


(EX1): Speaker adaptation

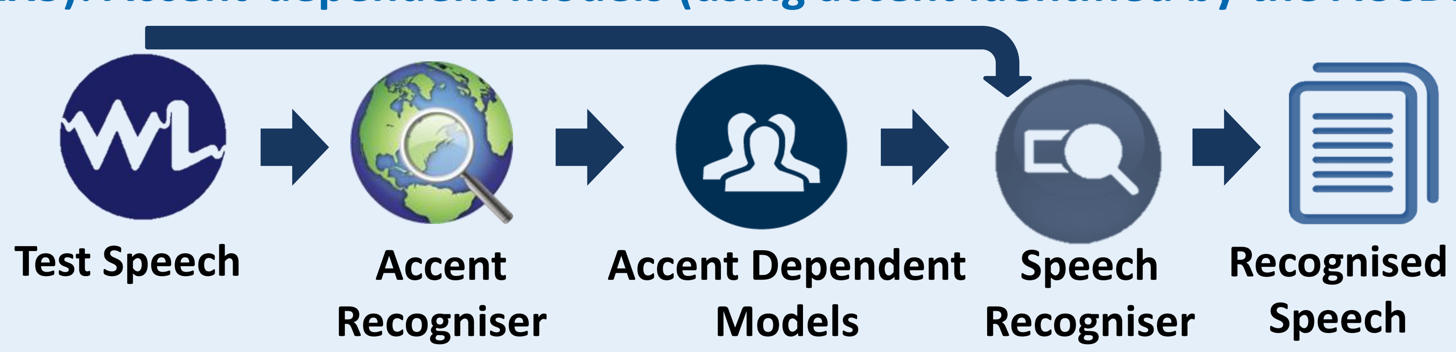


4. Methodology

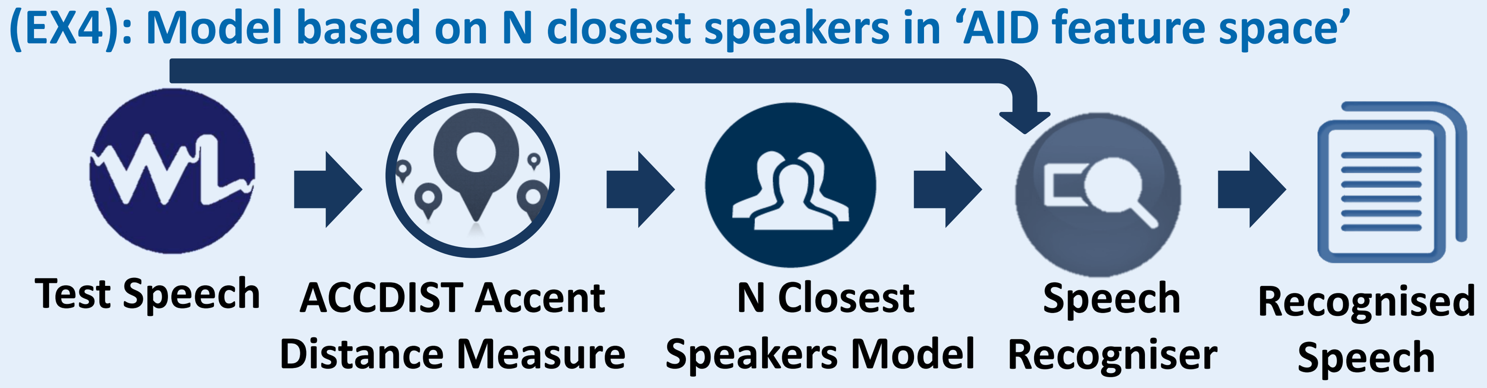
(EX2): Accent-dependent models (using prior knowledge of accent)



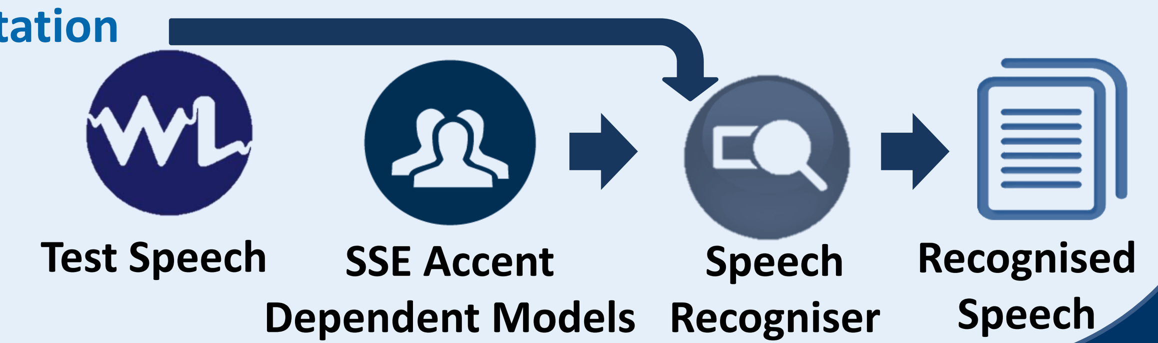
(EX3): Accent-dependent models (using accent identified by the ACCDIST)



(EX4): Model based on N closest speakers in 'AID feature space'



(EX5): SSE adaptation



5. Results

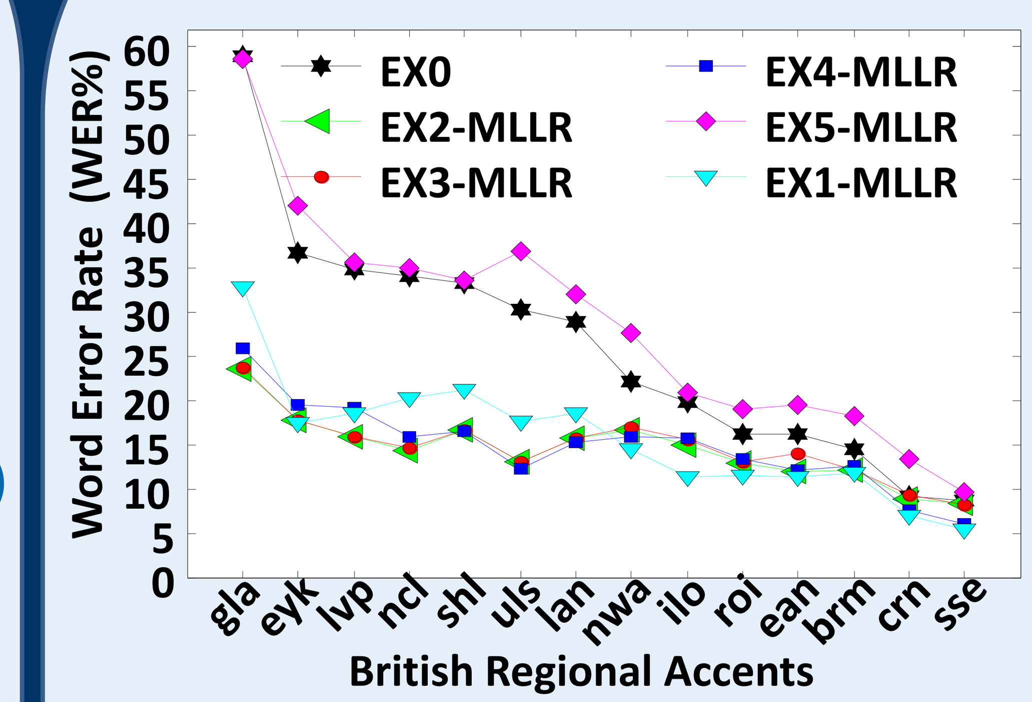


Figure2. Comparison of MLLR adaptation results for different methods

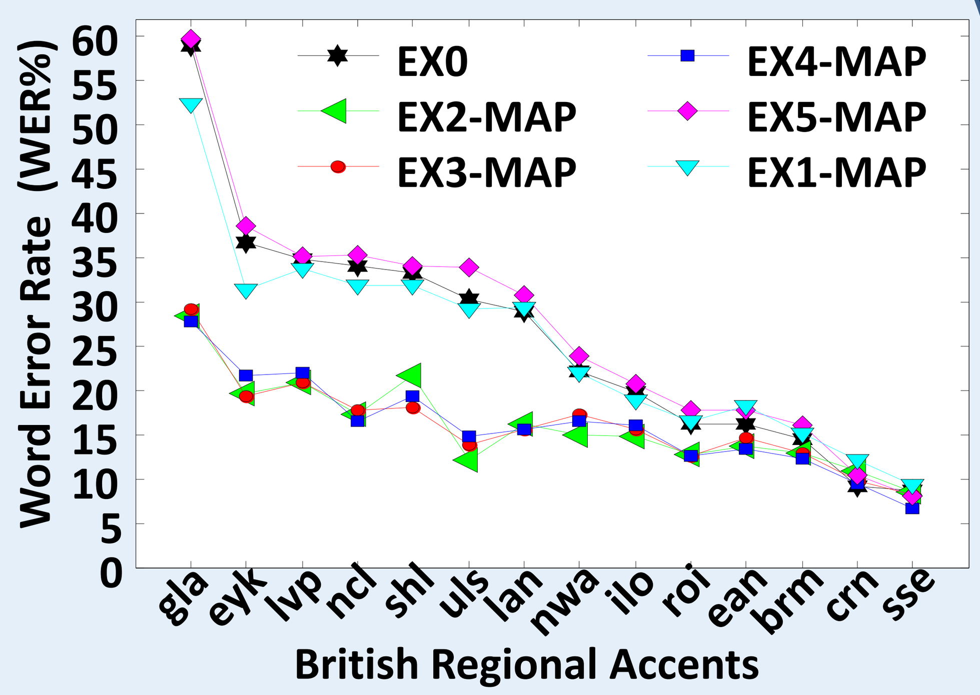


Figure3. Comparison of MAP adaptation results for different methods

EXP	Adaptation Method	MAP (WER%)	MLLR (WER%)
EX0	None	26.0	26.0
EX1	Speaker	25.5	15.9
EX2	True Accent	16.6	14.7
EX3	AID Accent	16.1	14.8
EX4	9 Nearest	16.4	15.6
EX5	SSE	27.3	28.7

Table1. Results summary

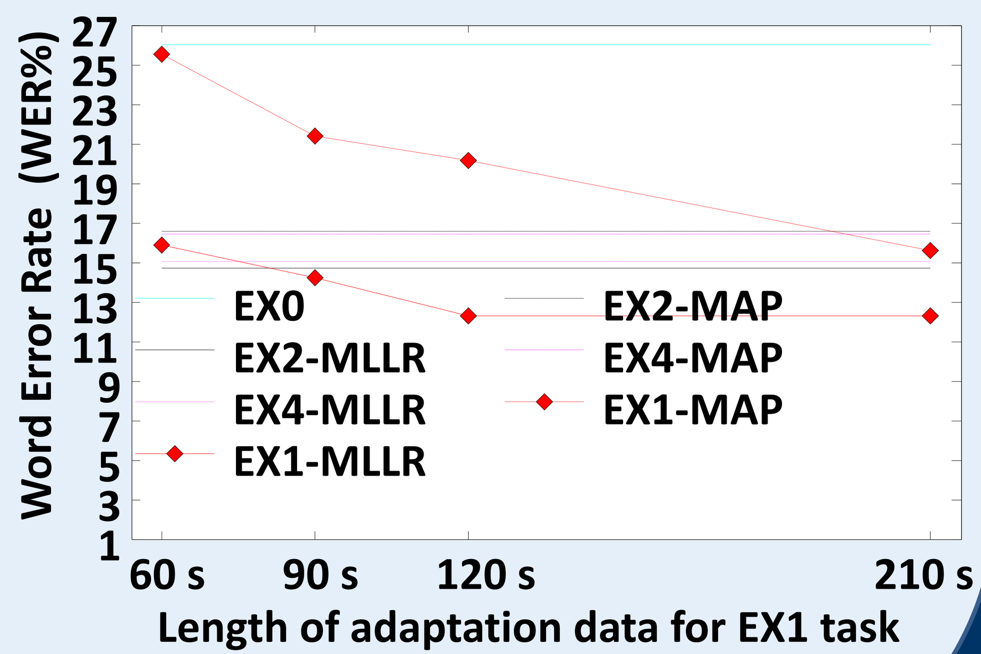


Figure4. Comparison of Speaker and accent adaptation results

6. Conclusions

As shown in Figures 2 and 3, methods EX2 to EX4 give similar performance, which is significantly better than the performance obtained with the baseline, accent-independent model (EX0). Results in Table 1 show relative reductions in ASR error rate of 37% and 44% for accent-dependent models built using MAP and MLLR adaptation respectively, compared with the baseline system (EX0).

According to Figure 4, using the 60 s of speech to identify an appropriate accent-dependent model outperforms using the same 60 s of speech for speaker-adaptation, by 35.8% and 7.6% for MAP and MLLR-based speaker adaptation.

All in all, we managed to use the accent-dependent acoustic modeling to develop both rapid and accent robust ASR system.

7. References

[1] Najafian, M., et al (2013) "Modelling Regional Accent for Automatic Speech Recognition" Submitted to Interspeech 2013.

[2] Huckvale, M., 2007. ACCDIST: an accent similarity metric for accent recognition and diagnosis. In: Müller, C. (Ed.), Speaker Classification II. Springer-Verlag, Berlin/Heidelberg, Germany, pp. 258–275.