The 2015 Sheffield System for Transcription of Multi–Genre Broadcast Media

O. Saz, M. Doulaty, S. Deena, R. Milner, R.W.M. Ng, M. Hasan, Y. Liu, T. Hain [1]

Speech and Hearing Group, Department of Computer Science, University of Sheffield, UK

{o.saztorralba, m.doulaty, s.deena, r.milner, wm.ng, m.hasan, y.liu, t.hain}@sheffield.ac.uk

Introduction

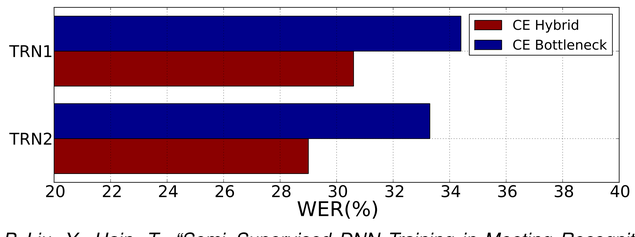
- Presentation of the University of Sheffield system for Task 1 of the MGB challenge
  - **Speech–to–text transcription of broadcast television**
- Four topics of research
  - Data selection and amount of training data
  - Automatic speech segmentation with DNNs and resegmentation
  - Language model adaptation to multiple genres
  - Modelling and adaptation to diverse acoustic domains and background conditions
- Final system is available for research in our **web–based recogniser http://www.webasr.org**

Basic system description

- The data were BBC broadcasts and subtitles **officially distributed on the MGB challenge**
  - Acoustic training data: 2,193 shows with 1,580 hours of audio and lightly supervised transcripts
  - Language training data: 10M words from 2,193 training shows and 648M words from historical subtitles
  - Development data: 47 shows with 28 hours of audio
  - 8 genres: **Advice, children's, comedy, competition, documentary, drama, events and news**
- Baseline systems were built with the following characteristics:
  - Acoustic models:
    - **Hybrid** DNN–HMM systems, from a 6–layer network with 10 contiguous PLP frames transformed by bMMI and CMLLR as inputs and 6,000 output targets. Both CE and sMBR target functions were used.
    - **Bottleneck** DNN–GMM–HMM system, with 39 PLP and 26 bottleneck features from a 4–layer DNN as input and 16 gaussians for 8,000 ML–trained triphone states. Both CE and sMBR target functions were used.
  - Lexicon: 50,000 usual words from the 2,193 training shows with pronunciation probabilities
  - Language models: Interpolated 4–gram from the 2 linguistic sources available

Data selection and training

- Transcription of the training data originated from subtitles with different quality. Two data selection techniques were studied
  - **TRN1**: 512 hours of lowest Word Matching Error Rate (WMER) given on the training data. WMER is the error between the aligned subtitles and a lightly supervised decoding provided for the MGB challenge.
  - **TRN2**: 698 hours of highest posterior–based confidence measure. The confidence measure is an average of the posteriors for the monophones in the word/utterance obtained from a 4–layer DNN with 144 monophone states as targets [a].
- Recognition systems trained on extra 200 hours of data reduce WER by 1.5% absolute.



[a] *Zhang, P. Liu, Y., Hain, T., "Semi–Supervised DNN Training in Meeting Recognition",* SLT 2014*, Lake Tahoe, NV.*

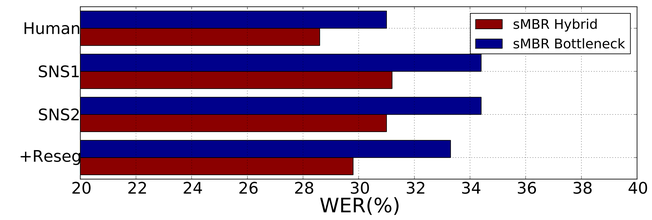
[1] *Supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology)*

Automatic segmentation

- Proposed system for automatic speech segmentation in two stages[b]
  - First stage: **2–layer DNN speech/nonspeech detection** with two possible training sets
    - **SNS1**: 1,552 hours (759 h. of speech, 793 h. of non–speech)
    - **SNS2**: 479 hours (363 h. of speech, 116 h. of non–speech)
  - Second stage: **Resegmentation of speech** based on converting to silence the lowest confidence[a] hypothesis words from an initial ASR stage
- Segmentation results on the dev set show that resegmentation after $SNS2$ DNN segmentation gives similar performance to $SNS1$ DNN segmentation with more balanced errors

	Speech time	Segments	Missed speech	False speech	Segmentation error
Human	19.5h.	30,702	0.0%	0.0%	**0.0%**
$SNS1$	18.3h.	17,713	6.6%	2.6%	**9.2%**
$SNS2$	21.8h.	15,337	1.3%	15.4%	**16.7%**
+Reseg	19.3h.	16,327	4.0%	5.4%	**9.4%**

- Recognition results show that resegmentation recovers more than 1% WER compared to initial DNN segmentation ($SNS1$ and $SNS2$), but there is still a 2% WER degradation compared to human segmentation



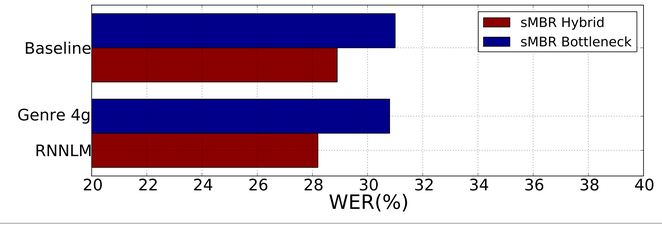
[b] *Milner, R., Saz, O., Deena, S., Doulaty, M., Ng, R.W.M., Hain, T., "The 2015 Sheffield system for longitudinal diarisation of broadcast media",* ASRU 2015*, Scottsdale, AZ*

Multi–genre language modelling

- Adaptation to linguistic diversity of each broadcast genre was proposed in this way.
  - An LDA model is trained on the genre–tagged 10M–word subtitles, to **describe each document as a set of hidden topics**, the distribution of hidden topics from the LDA are used as features to train SVMs for each of the 8 genres using.
  - LDA topic distribution is calculated for the 650M–word subtitles and genres are estimated
  - Genre–dependent **4–grams and RNNLMs** are trained and used for lattice rescoring
- Genre–dependent 4–grams reduce perplexity, with RNNLMs giving further reduction

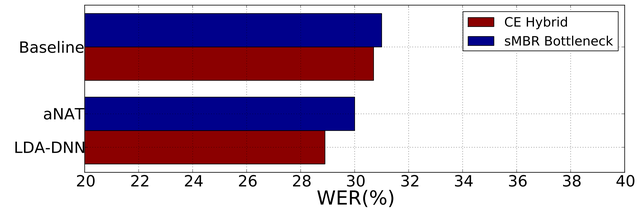
LM	Adv.	Child.	Comed.	Compet.	Docum.	Dram.	Even.	News
Baseline 4–gram	94.5	101.4	102.1	104.2	129.4	83.9	126.3	137.1
Adapted 4–gram	87.2	92.1	93.8	94.5	124.1	78.4	120.0	125.1
Adapted RNNLM	58.6	62.7	59.6	50.5	68.7	60.4	64.0	67.2

- In recognition, adapted 4–grams produced 0.2% WER reduction in $Bottleneck$ configuration, and genre–based RNNLMs 0.7% WER reduction in $Hybrid$ configuration.



Acoustic background modelling

- Two strategies for background and domain compensation in multi–genre broadcasts
  - DNN adaptation using one–hot auxiliary input features based on 64 automatically derived **acoustic domains extracted using LDA**[c]
  - GMM–HMM asynchronous adaptation and Noise Adaptive Training using **8 asynchronous noise CMLLR transformations**[d]
- WER reduction is 1% for the asynchronous NAT $Bottleneck$ system and 2% for the domain adapted $Hybrid$ system

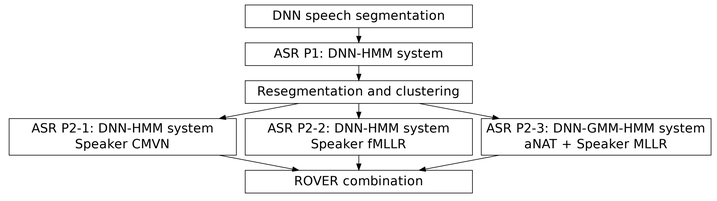


[c] *Doulaty, M., Saz, O., Hain, T., "Latent Dirichlet Allocation Based Organisation of Broadcast Media Archives for Deep Neural Network Adaptation",* ASRU 2015*, Scottsdale, AZ.*

[d] *Saz, O., Hain, T., "Asynchronous Factorisation of Speaker and Background with Feature Transforms in Speech Recognition",* Interspeech 2013*, Lyon, France.*

Final system

- Final system features:
  - Acoustic models trained on $TRN2$ 700 hours of training data
  - Segmentation based on DNN and resegmentation
  - 3 complementary systems for system combination



- Results on the dev set show a 4% absolute WER reduction from the initial ASR to the final combination, with a larger **improvement in Children's, Comedy and Events shows**.

System	Adv.	Child.	Comed.	Compet.	Docum.	Dram.	Even.	News	Global
ASR P1	23.1%	36.5%	45.4%	25.1%	30.0%	40.8%	36.4%	14.1%	**31.2%**
ASR P2-1	22.8%	31.0%	42.9%	24.1%	28.4%	38.6%	33.6%	14.2%	**29.4%**
ASR P2-2	23.0%	31.2%	42.8%	24.2%	28.5%	39.0%	33.5%	13.8%	**29.4%**
ASR P2-3	23.7%	32.0%	45.3%	25.1%	29.3%	40.5%	34.3%	15.0%	**30.5%**
ROVER	21.6%	27.7%	40.9%	22.7%	26.6%	37.1%	31.3%	13.2%	**27.5%**

Conclusions

- Work focused on domain adaptation, both on the acoustic and the linguistic side of the systems, with 1-2% reduction in errors for different individual techniques
- Improvement in automatic segmentation recovers more than 1% of errors over a DNN system
- Extra work is required in data selection for AM training, due to the unreliability of the subtitles