Revisiting Joint Decoding based Multi-talker Speech Recognition with DNN Acoustic model

Martin Kocour Kateřina Žmolíková Lucas Ondel Ján Švec
Marc Delcroix Tsubasa Ochiai Lukáš Burget Jan Černocký

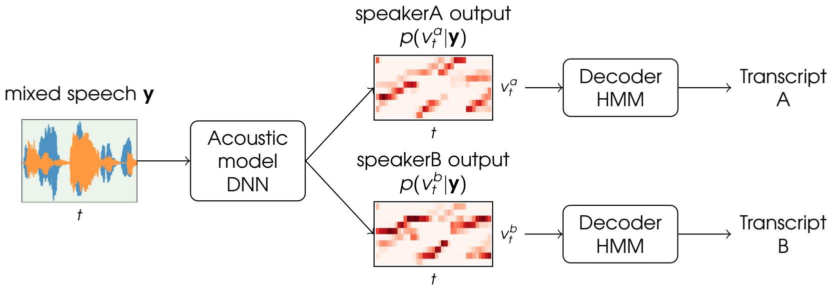
Brno University of Technology, Czechia NTT corporation, Japan

ikocour@fit.vutbr.cz

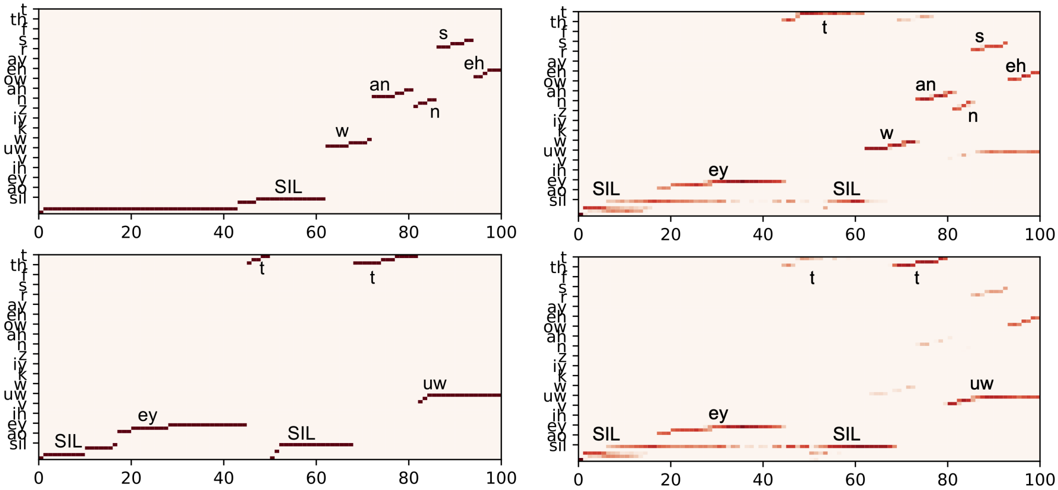
BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

September 22, 2022

Previous Works: Multi-talker ASR with separate decoding



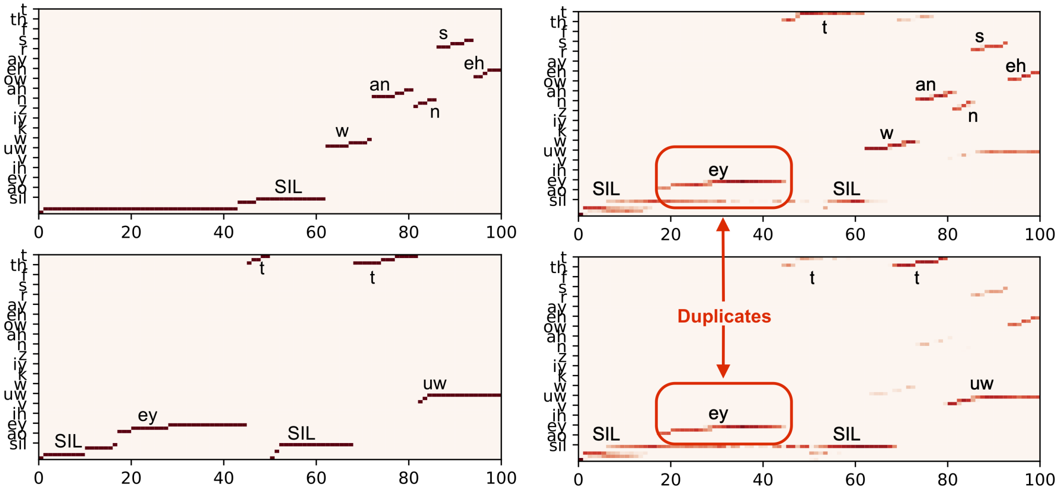
- The acoustic model is trained to produce pdf-posteriors for each speaker separately.
- The decoding is performed independently.
- Not optimal especially for mixtures with similarly sounding speakers (e.g. same-gender speakers).



(a) Source phone alignments

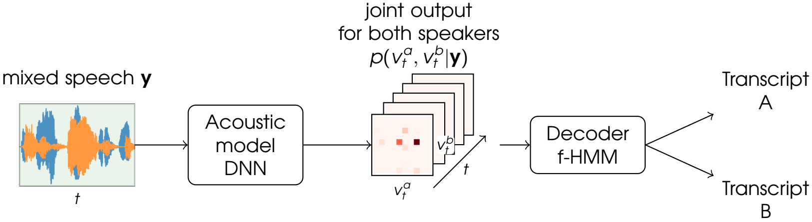
(b) Separate phonetic posteriorgrams

Example I: Mixture with 2 speakers of the same gender



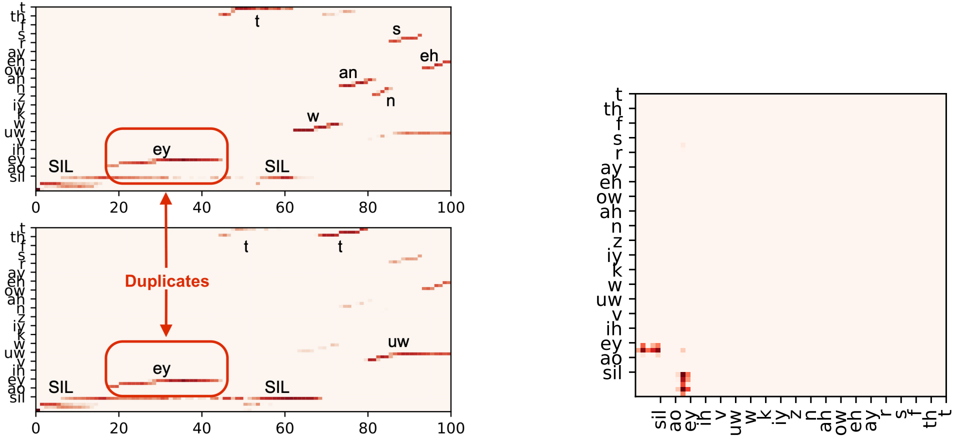
(a) Source phone alignments (b) Separate phonetic posteriorgrams

- Acoustic model assigns high probability of phone 'ey' to both speakers.



Main idea:

- Extend factorial HMM model to DNNs.
- The decoder also considers the other speaker's speech states.
- This can improve the performance especially in cases where acoustic model is unable to separate the speech.

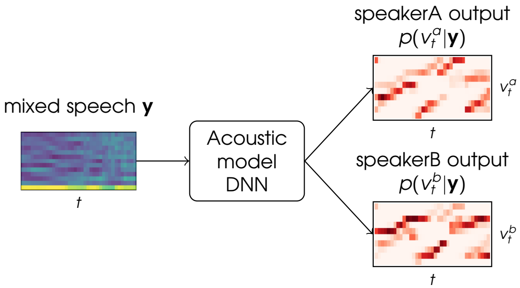


(a) Separate phonetic posteriorgrams

(b) Joint posteriors

- The proposed AM, which produces the joint posteriors, is able to distinguish that the phone ey comes from either first or second speaker, but not from both speakers

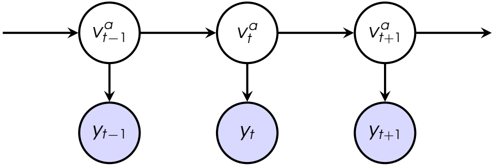
Conventional approaches for Multi-talker Speech Recognition



- The acoustic model predicts the posterior probabilities for each speaker k separately:

$$p(v_t^a|y_t), \;\; p(v_t^b|y_t) = f_{\mathrm{NN}}(y_t) \tag{1}$$

- The acoustic model is forced to learn how to separate the mixed speech.
- The model is trained in permutation-invariant fashion.



- Recognising speech involves finding the most likely state sequence $\hat{\mathbf{v}}^a$ given observed data $\mathbf{y}$ (i.e. MAP state sequence).
- The MAP state sequence is obtained by Viterbi algorithm, where messages $m(v_t)$ are defined as:

$$m(v^a_{t+1}) = \max_{v^a_t} p(v^a_{t+1}|v^a_t)m(v^a_t)\bar{p}(y_t|v^a_t) \tag{2}$$

$$\tilde{v}_t(v^a_{t+1}) = \arg\max_{v^a_t} p(v^a_{t+1}|v^a_t)m(v^a_t)\bar{p}(y_t|v^a_t). \tag{3}$$

- The MAP state sequence $\hat{\mathbf{v}} = [\hat{v}_1, \ldots, \hat{v}_T]$ is recovered by backtracking:

$$\hat{v}_t = \tilde{v}_t(\hat{v}_{t+1}) \tag{4}$$

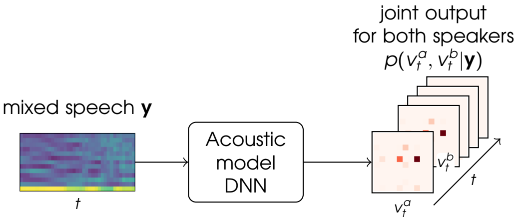
where $\hat{v}_T = \arg\max_{v_T} m(v_T)$ initiates the recursion.

- Conventional ASR models assume conditional independence of the state sequences of the speakers given the observation, i.e.

$$p(\mathbf{v}^a, \mathbf{v}^b|\mathbf{y}) = p(\mathbf{v}^a|\mathbf{y})p(\mathbf{v}^b|\mathbf{y}). \tag{5}$$

- We expect the acoustic model to fully solve the "separation" of the speakers, i.e., fully attribute parts of the mixed speech signal to the different outputs of the network.
- This may be challenging, especially when the speakers' voices are very similar.
- There is no interaction between the decoders of the individual speakers.
- This can lead to duplicity where the same phoneme or word is attributed to both speakers.

Proposed Multi-talker ASR with joint decoding

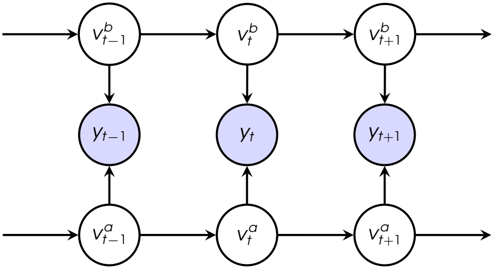


- To take account of the dependencies, we model the joint probability of mixed speech $\mathbf{y}$ and hidden state sequences $\mathbf{v}^a, \mathbf{v}^b$ in Factorial HMM framework:

$$p(\mathbf{y}, \mathbf{v}^a, \mathbf{v}^b) = \prod_t p(y_t|v_t^a, v_t^b)p(v_t^a|v_{t-1}^a)p(v_t^b|v_{t-1}^b), \tag{6}$$

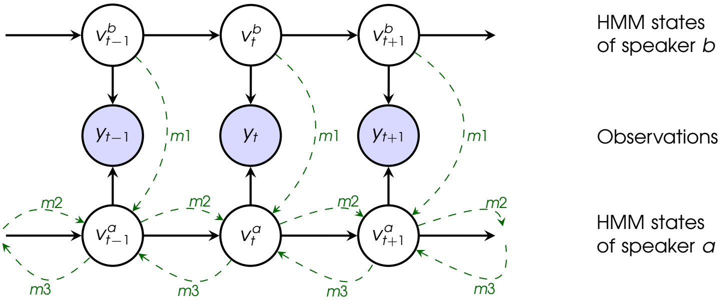
- where $p(y_t|v_t^a, v_t^b)$ is derived from a neural network that predicts the posterior probabilities of a tuple of states (v_t^a, v_t^b) as

$$p(v_t^a, v_t^b|y_t) = g_{\mathrm{NN}}(y_t), \tag{7}$$



- Standard Viterbi algorithm can be used to exactly infer the MAP hidden state sequences $\hat{\mathbf{v}}^a, \hat{\mathbf{v}}^b$ in FHMM.
- We need to compute messages between all possible combinations of hidden states for all speakers, i.e. decoding network is a Cartesian product of HMM states from all speakers.
- Its time complexity is $O(TKV^{K+1})$ for K speakers, HMM with V states and utterance time T.
- It scales exponentially w.r.t. to a number of speakers.

(Rennie S., Hershey J., and Olsen P., *Single-Channel Multitalker Speech Recognition*, IEEE Signal Processing Magazine)

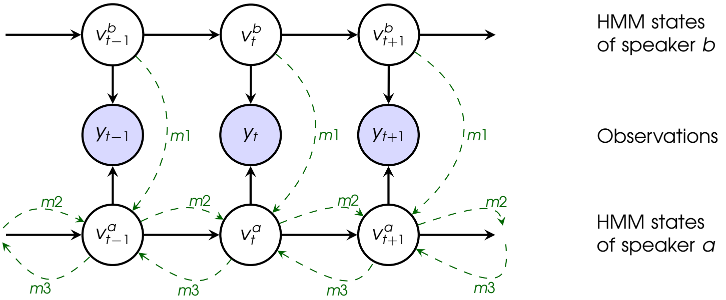


- The messages are passed between variables, which share the common factors according to a predefined schedule.
- Our message passing schedule has the following form for one speaker:

$$\tilde{p}(y_t|v^a_t) = \max_{v^b_t} \bar{p}(y_t|v^a_t, v^b_t)\tilde{p}_{fw}(v^b_t)\tilde{p}_{bw}(v^b_t) \qquad \text{(m1)}$$

$$\tilde{p}_{fw}(v^a_t) = \max_{v^a_{t-1}} p(v^a_t|v^a_{t-1})\tilde{p}_{fw}(v^a_{t-1})\tilde{p}(y_{t-1}|v^a_{t-1}) \qquad \text{(m2)}$$

$$\tilde{p}_{bw}(v^a_t) = \max_{v^a_{t+1}} p(v^a_{t+1}|v^a_t)\tilde{p}_{bw}(v^a_{t+1})\tilde{p}(y_{t+1}|v^a_{t+1}) \qquad \text{(m3)}$$



- When all messages for the first speaker are computed, the process is repeated for the next speaker, while messages of other speakers are fixed.
- The maximizing arguments $\tilde{v}_t$ for all messages are also stored to recover the MAP state sequences in a manner analogous to the Viterbi algorithm.
- The whole process is repeated until it converges or some number of iterations is reached.
- The time complexity of the proposed LBP inference is $O(TKV^2)$, and thus it scales linearly w.r.t. to a number of speakers.

Experiments

Comparing MT-ASR with separate and proposed joint decoding

- mixed TIDIGIT dataset, where each mixture consists of exactly 2 overlapping speakers
- Model architecutre is same except the final layers:
  - AM for separate decoding contains 2 output layers of size 62 each
  - AM for joint decoding contains 1 output layer of size 62×62
- hybrid ASR system, where decoding network is similar to the network used in Kaldi TIDIGIT recipe

	Arch	Output (dim)	#params	Separate	Joint	Kaldi
1	5L-TDNN	(62, 62)	1.9 M	26.09	-	25.99
2	5L-TDNN	(3844)	3.3 M	17.55^{Σ}	15.79	17.83^{Σ}
3	10L-TDNN	(62, 62)	4.1 M	18.68	-	18.66
4	10L-TDNN	(3844)	5.5 M	16.36^{Σ}	**14.70**	16.97^{Σ}

- For a more fair comparison, we include a third *separate-marginal* method marked with Σ-symbol, which combines the AM predicting joint posteriors with separate decoding

$$p(v_t^a|y_t) = \sum_{v_t^b} p(v_t^a, v_t^b|y_t) \tag{8}$$

$$p(v_t^b|y_t) = \sum_{v_t^a} p(v_t^a, v_t^b|y_t) \tag{9}$$

- This allows us to separately evaluate the benefit of the joint posterior output (which also induces an increased number of parameters) and the benefit of the joint decoding itself

- Comparison of the systems on mixtures containing speakers of same or different gender.

Genders	Separate (%WER)		Joint (%WER)
F + F	30.54	28.57^{Σ}	21.45
M + M	32.61	27.87^{Σ}	27.12
same	31.56	28.23^{Σ}	24.26
opposite	6.17	4.85^{Σ}	5.42

Summary and Future plans

Summary

- We proposed a new architecture for multi-talker speech recognition with joint decoding.
- It has the potential to improve performance in challenging conditions where it may be difficult to achieve high separation by simply relying only on the acoustic information.
- Joint decoder was implemented in Julia using new MarkovModels toolkit (implemented by Lucas Ondel and Martin Kocour).

Future plans

- Presented results are just a proof-of-concept experiments
- Extend the idea to larger task, e.g. WSJ0-2mix or even more realistic Chime6.
- Train the proposed AM with MMI/CTC loss, where joint pdf-posteriors would be approximated by LBP

Thank you four your attention!

Questions? ikocour@fit.vutbr.cz

- Performance measured on single-talker speech.

	Arch	Output (dim)	#params	Separate	Joint
1	10L-TDNN	(62, 62)	4.1 M	22.17	-
2	10L-TDNN	(3844)	5.5 M	21.35^{Σ}	**15.69**

1. 22.17 % WER (6338 / 28583, 3843 ins, 2407 del, 88 sub)
2. 21.35^{Σ} % WER (6103 / 28583, 3825 ins, 2119 del, 159 sub)
3. **15.69** % WER (4485 / 28583, 2729 ins, 1552 del, 204 sub)

Comparing MT-ASR with separate and proposed joint decoding

- Dataset: mixed TIDIGIT
  - overlapped speech of 2 speakers, where each is pronouncing some sequence of digits
  - train: 52.5 hours, valid: 5.3 hours of speech
- Acoustic model: CNN layers with batch normalization, and ReLu activation function
  - both PIT-ASR with separate decoding and the proposed ASR with joint decoding shares the same architecture
  - AM for separate decoding contains 2 output layers with size 62 each
  - AM for joint decoding contains 1 output layer with size $62 \times 62 = 3844$
  - PIT-CE objective function (frame-level alignments from Kaldi)
- Decoding network: similar to Kaldi TIDIGT setup for monophone GMM-HMM ASR system
  - unigram LM, where each word (i.e., digit) is equally likely
  - silence is modeled by 5 HMM states
  - other 19 phones are modeled by 3-state HMM
  - states do not share emission probabilities, i.e., we have 62 PDFs in total