SpeeD's DNN Approach to Romanian Speech Recognition

Alexandru-Lucian Georgescu, Horia Cucu and Corneliu Burileanu

Speech & Dialogue (SpeeD) Research Laboratory
University "Politehnica" of Bucharest (UPB)

SpeeD ASR Improvements

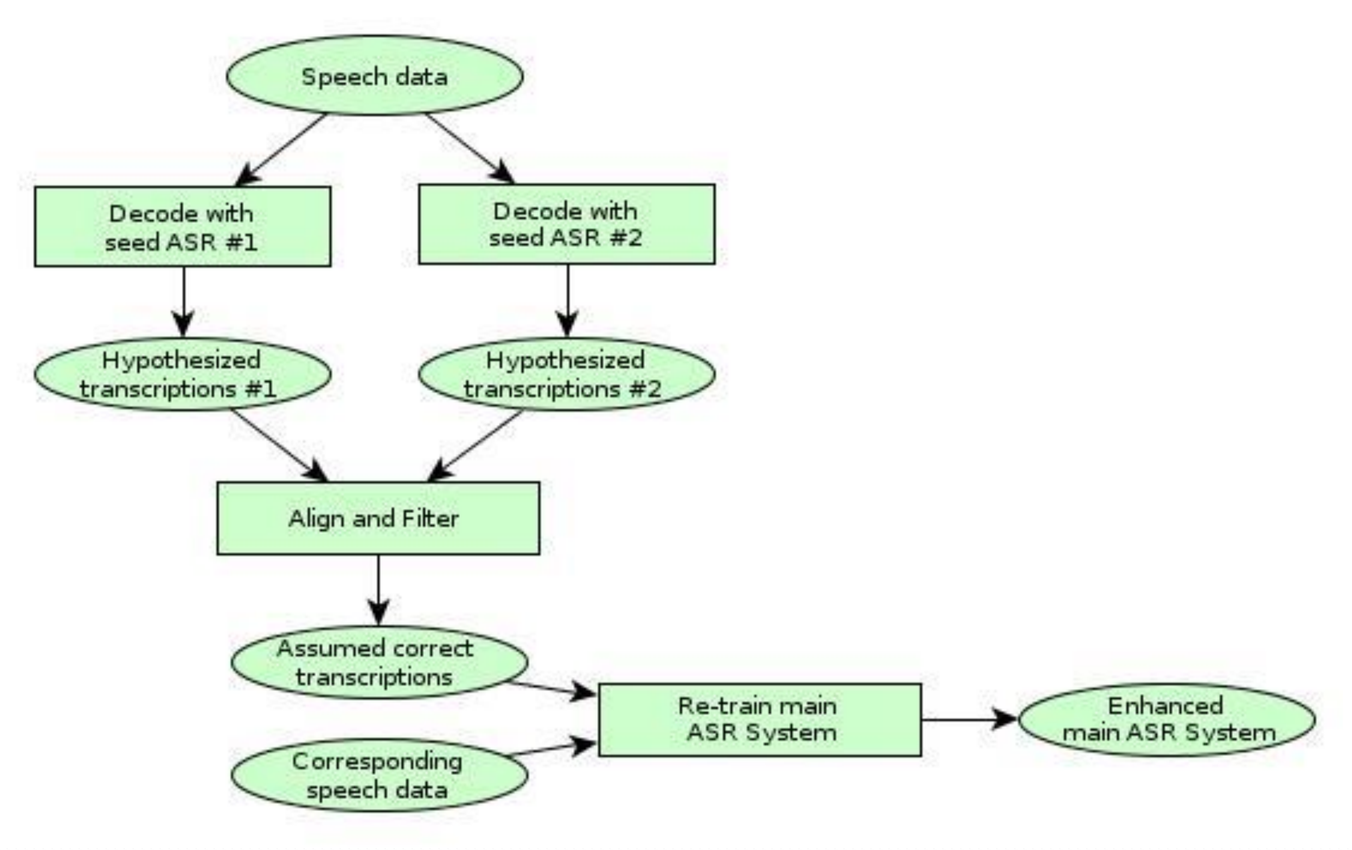
- SpeeD's 2014 LVCSR system [Cucu, 2014]
  - MFCCs or PNCCs used as speech features
  - HMM-GMM acoustic models trained on ~125 hrs of speech
  - 64k words 3-gram language models trained on ~200M word tokens

- SpeeD's LVCSR improvements since 2014
  - Speech and text resources acquisition
  - Improved language models: larger vocabulary, more grams
  - Improved GMM acoustic models and DNN acoustic models
  - Speech feature transforms (LDA, MLLT)
  - Lattice rescoring after speech decoding

Speech Corpora

- Read Speech Corpus (RSC) – train & eval
  - Created by recording various predefined texts
  - Voluntary speakers used an online recording platform
  - 106 hrs of read speech from 165 different speakers
- Spontaneous Speech Corpus (SSC) – train
  - Created using lightly supervised ASR training [Buzo, 2013]
    - broadcast news and talk shows + approximate transcriptions collected over the Internet
  - 27 hrs of speech
- Spontaneous Speech Corpus (SSC) – eval
  - Manually annotated to obtain 100% error-free corpus
  - 3.5 hrs of speech (2.2 hrs clean, 1.3 hrs degraded conditions)
- Spontaneous Speech Corpus 2 (SSC 2) - train
  - Unsupervised annotation methodology [Cucu, 2014]
  - 350 hrs of un-annotated broadcast news -> 103 hrs of annotated speech

Unsupervised Speech Corpus Extension



Improved Acoustic Models

- HMM – GMM framework
  - Discriminative training: Maximum Mutual Information (MMI) [Povey, 2008]
    - Maximizes the posterior probability for the training utterances
  - Speaker Adaptive Training (SAT) [Povey, 2008]
    - Adapts acoustic model to speaker characteristics (if speaker info is available)
  - Algorithms available in Kaldi ASR toolkit

- DNN framework
  - Time Delay Neural Network (TDNN) [Zhang, 2014] [Peddinti, 2015]
    - Able to learn long-term temporal dependencies
    - Input: 9 frames of speech
  - Speech features: standard MFCCs + iVectors (useful for speaker adaptation)
  - Input layer size: couple of thousand neurons
  - Output layer size: couple of hundred neurons
  - Hidden layers: 3 - 6 hidden layers with around 1200 neurons
  - Framework and algorithms available in Kaldi ASR toolkit

Improved Language Models

- Kaldi ASR toolkit allows using LMs with larger vocabularies than CMU Sphinx ASR toolkit (limited at 64k words)

- Text corpora used for language modeling
  - Extended by collecting new texts from the Internet
    - 169M word tokens (in 2014) -> 315M word tokens (in 2017)
  - Text collected from the Internet needed diacritics restoration [Petrica, 2014]
  - Talk shows transcriptions (40M word tokens) already available

- Language Models (LMs)
  - Statistical n-gram models
  - Created with SRI-LM by interpolating text corpora with various weights
  - Various n-gram orders: from 1-gram to 5-gram
  - Various vocabulary sizes: 64k, 100k, 150k and 200k words

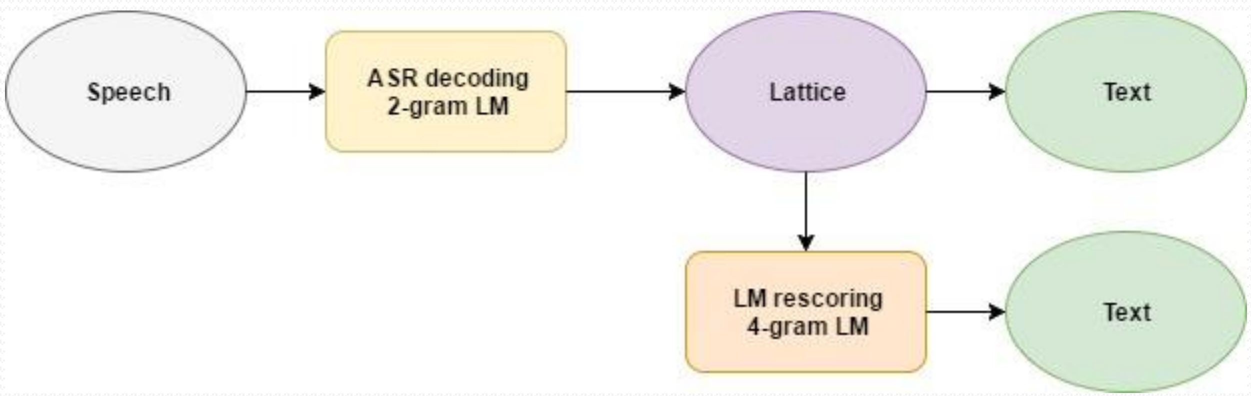
Lattice rescoring

- After ASR decoding with short history LM (2-gram):

  aceste este un peste de recunoaștere automată a vorbi ei

- After LM rescoring with longer history LM (4-gram):

  acesta este un test de recunoaștere automată a vorbirii



Lattice rescoring concept

Experimental setup. Speech Corpora

- Read Speech Corpus (RSC)
  - read speech utterances in silent environment
  - clean speech
- Spontaneous Speech Corpus (SSC)
  - spontaneous utterances from talk shows and news broadcasts
  - clean and spontaneous speech, sometimes affected by background noise

Purpose	Set	Size	
Training	RSC-train	94 h , 46 m	225 h, 31 m
	SSC-train 1	27 h, 27 m	
	SSC-train 2	103 h, 17 m	
Evaluation	RSC-eval	5 h, 29 m	8 h, 58 m
	SSC-eval	3 h, 29 m	

Experimental setup. Speech features

- Mel-frequency cepstral coefficients (MFCCs)
- Extracted from 25 ms signal window length, shifted by 10 ms
- Final feature vector: 13 MFCCs x 9 frames
- Features transforms
  - Cepstral Mean and Variance Normalization (CMVN)
    - Normalize the mean and variance of raw cepstra
    - Eliminate inter-speaker and environment variations
  - Linear Discriminant Analysis (LDA)
    - Reduce features space dimension keeping class discriminatory information
  - Maximum Linear Likelihood Tranform (MLLT)
    - Capture correlation between the feature vector components

Experimental setup. Acoustic Models

- HMM – GMM framework
  - Speech features: 13 MFCCs + Δ + ΔΔ
  - LDA + MLLT
  - 2.500 – 5.000 senones, 30.000 – 100.000 Gaussian Densities
  - Maximum Mutual Information (MMI)
    - Maximize the posterior probability for the training utterances
  - Speaker Adaptive Training (SAT)
    - Adapt acoustic model to speaker characteristics

- Time Delay Neural Network (TDNN)
  - Speech features: 40 MFCCs x 9 frames + 1 iVector of 100 elements
  - LDA + MLLT
  - Input layer size: 3500 and 4400 neurons
  - Output layer size: 350 and 440 neurons
  - 3 and 6 hidden layers
  - Up to 15 training epochs

Experimental setup. Language Models

- Text corpora used for language modeling
  - Collected news from the Internet (315 M word tokens)
  - Broadcasted talk shows (40M word tokens)

- Language Models (LMs)
  - Statistical n-gram models
  - Created with SRI-LM by interpolating text corpora with 0.5 weight
  - Different n-gram order: from 1-gram to 5-gram
  - Different vocabulary size: 64k, 100k, 150k and 200k words

Experimental results

- HMM –GMM framework
- LM used: 3-gram, 64k words

Acoustic model		Feat. Transf. & training tech.	WER [%]	
#Senones	# Gaussians		RSC-eval	SSC-eval
2.500	30.000	n/a	12.3	29.7
4.000	50.000	LDA+MLLT	11.3	28.9
5.000	100.000	+SAT	9.7	27.5
5.000	100.000	+MMI	**9.0**	**26.4**

Experimental results

- DNN framework
- DNN configurations
  - 3500 in. neurons, 350 out. neurons, 6 hidden layers, 8 epochs
  - 4400 in. neurons, 440 out. neurons, 6 hidden layers, 8 epochs
  - 4400 in. neurons, 440 out. neurons, 3 hidden layers, 15 epochs
- LM used: 3-gram, 64k words

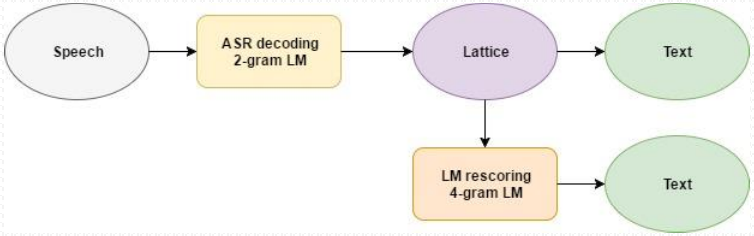
DNN Config.	# train. Epochs	WER [%]	
		RSC-eval	SSC-eval
3500 in neurons 350 out neurons 6 hidden layers	1	6.4	21.7
	2	**6.2**	**21.0**
	3	6.3	20.7
	4	6.4	21.0
	5	6.4	21.2
	8	6.9	22.1

Languge models evaluation

Vocabulary size	ASR decoding LM order		WER [%] RSC-eval w/o LM rescoring	WER [%] SSC-eval w/o LM rescoring
100 k words	1-gram		15.0	36.5
	2-gram		6.44	23.4
	3-gram		5.18	20.6
150 k words	1-gram		14.6	36.4
	2-gram		6.26	23.3
	3-gram		5.00	20.5
200 k words	1-gram		14.2	36.4
	2-gram		**5.90**	**23.2**
	3-gram		**4.62**	**20.5**

Lattice rescoring

Vocabulary size	ASR decoding LM order		WER [%]			WER [%]	
			RSC-eval	SSC-eval		RSC-eval	SSC-eval
			w/o LM rescoring			with LM rescoring	
100 k words	1-gram		15.0	36.5		6.06	22.5
	2-gram		6.44	23.4		5.04	20.3
	3-gram		5.18	20.6		5.05	20.1
150 k words	1-gram		14.6	36.4		5.81	22.4
	2-gram		6.26	23.3		4.85	20.3
	3-gram		5.00	20.5		4.85	20.1
200 k words	1-gram		14.2	36.4		5.39	22.4
	2-gram		**5.90**	**23.2**		**4.49**	**20.2**
	3-gram		**4.62**	**20.5**		**4.48**	**20.0**



Memory consumption. Real time factor

- Intel Xeon 3.2 GHz with 16 cores
- 192 GB RAM

LM order	Decoding max memory	Decoding time [xRT]	
		RSC-eval	SSC-eval
1-gram	~ 1.5 GB	0.04	0.08
2-gram	**~ 8.5 GB**	**0.05**	**0.08**
3-gram	~ 30 GB	0.06	0.10

Overall improvement

SpeeD LVCSR System		WER [%]	
Acoustic model	**Language Model**	**RSC-eval**	**SSC-eval**
HMM – GMM (CMU Sphinx, 2014)	64 k words, 3-gram	14.8	39.1
HMM – GMM (CMU Sphinx, 2017)	64 k words, 3-gram	12.6	32.3
HMM – GMM (Kaldi, 2017)	64 k words, 3-gram	9.0	26.4
DNN (Kaldi, 2017)	64 k words, 3-gram	6.2	21.0
	200 k words, 2-gram (dec), 4-gram(resc)	**4.5**	**20.2**

Conclusions

- Several improvements of SpeeD LVCSR system for Romanian language were presented
- The application of feature transforms, discriminative training and speaker adaptive training algorithms led to a lower WER in HMM-GMM acoustic models
- The use of DNN acoustic models is the most important change
  - Relative WER improvements between **20.7%** to **30.8%** over HMM–GMM models
- Increasing the LM size & the use of lattice rescoring triggered a lower WER
- The overall relative WER improvement over the 2014 system
  - **70%** on read speech
  - **48%** on spontaneous speech

Thank you!