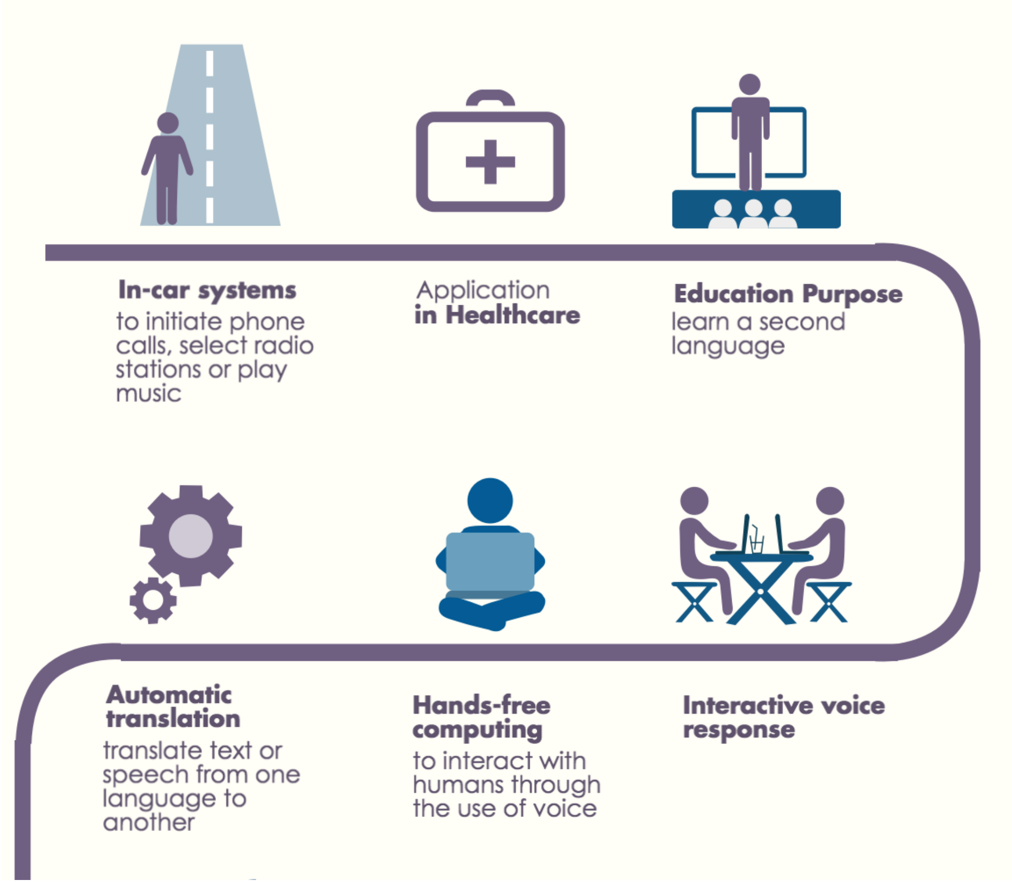
Speech Recognition and Accent Classification

Ian Lawson, Gazi Naven, Tolga Aktas

Motivation



In-car systems
to initiate phone calls, select radio stations or play music
Application in Healthcare
Education Purpose
learn a second language
Automatic translation
translate text or speech from one language to another
Hands-free computing
to interact with humans through the use of voice
Interactive voice response

Outline

- Early Stages of Automated Speech Recognition
- Advent of Deep Neural Nets in Speech Recognition
- Automated Accent Recognition

Roadmap of Speech Recognition

- 1970s: CMU spearheads the “Speech Understanding Research” (SUR), sponsored by DARPA.
  - Harpy recognizes ~1000 words.
  - ASR is highly use-specific, not very generalizable.
- 1980s & 1990s: Things get statistical
  - Hidden Markov Models (HMM)
    - Phones as Latent Variables
    - Audio features (e.g. MFCC) as Observations
  - Stochastic Language & Acoustic Modelling
  - Training GMMs for acoustic modelling, the distribution of features for phones.
- 2000 - present:
  - Neural Networks to replace GMM in HMM
  - HMM - free models
      - Sequence-to-sequence modelling
      - Abundant data

Speech Recognition Before DNN

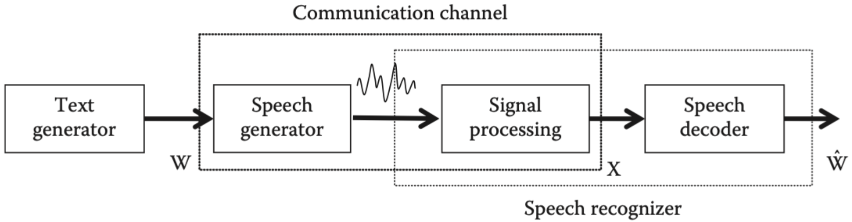


FIGURE 15.1 A source-channel model for a typical speech-recognition system.

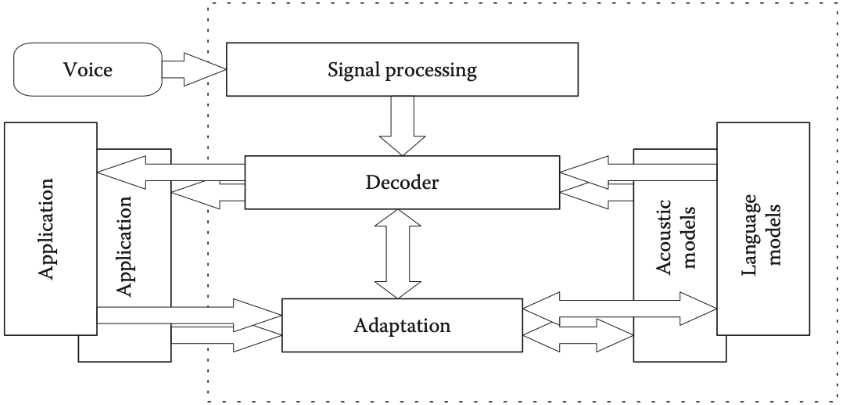


FIGURE 15.2 Basic system architecture of a speech-recognition system.

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{w}} P(\mathbf{W}|\mathbf{A}) = \arg\max_{\mathbf{w}} \frac{P(\mathbf{W})P(\mathbf{A}|\mathbf{W})}{P(\mathbf{A})}$$

Acoustic Models

- Statistical representation of feature vectors from waveforms
- Modeling via
  - HMMs (Most common)
  - (super) segment models
  - Conditional random fields
  - Maximum entropy models

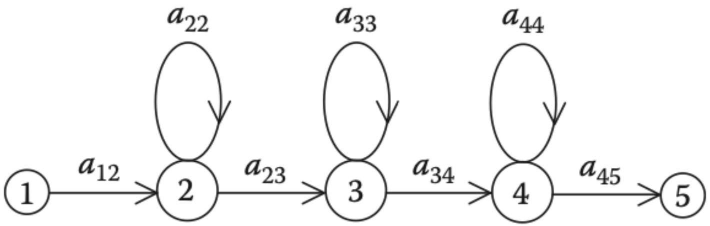


FIGURE 15.3 Illustration of a five-state left-to-right HMM. It has two non-emitting states and three emitting states. For each emitting state, the HMM is only allowed to remain at the same state or move to the next state.

[4] [5]

Language Models

- Reflect how frequently a word, **W,** occurs in a sentence

$$
\begin{aligned}
P(\mathbf{W}) &= P(w_1, w_2, \ldots, w_n) \\
&= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \cdots P(w_n|w_1, w_2, \ldots, w_{n-1}) \\
&= \prod_{i=1}^{n} P(w_i|w_1, w_2, \ldots, w_{i-1})
\end{aligned}
$$

where $P(w_i|w_1, w_2, \ldots, w_{i-1})$ is the probability that w_i will follow given that the word sequence $w_1, w_2, \ldots, w_{i-1}$ was presented previously

[6] [7]

What's going on ?

- As of now, the research is focused on the interaction of three components:
    - Language model : A probabilistic model of word sequences, modelling $P(W_{t+1} | W_t)$.
    - Pronunciation model: Mapping words to phonemes/graphemes/<a representative building block>
    - Acoustic model: Mapping from phonemes/graphemes/<a representative building block> to audio features
- What has changed?
    - Pre-HMM
    - HMM - GMM
    - HMM - **DNN: Neural Nets replace GMM for computing acoustic models.**
    - Seq2Seq models: LM, PM, AM are jointly modelled instead of separate models.
- Research is on finding better acoustic models, better representations

Deep Neural Networks for Acoustic Modeling in Speech Recognition

G. Hinton et. al.

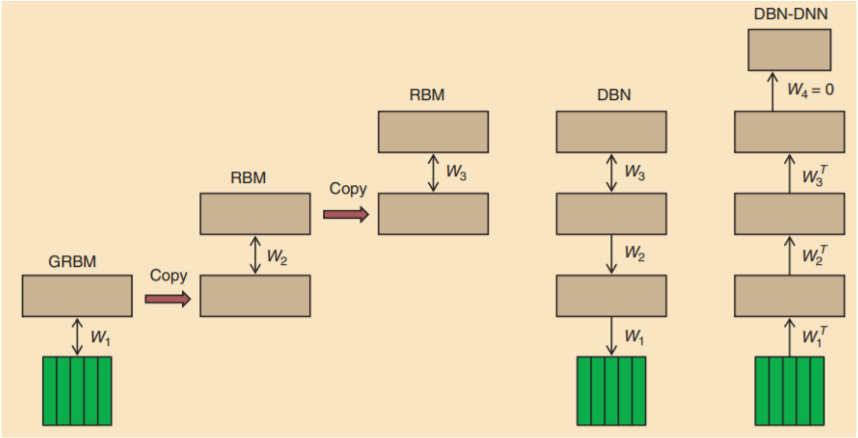
DNN - HMM (Hinton et. al)

- Previously:
    - GMM - HMM models
    - GMM: Learn the distribution for acoustic modelling, (which variable is responsible for each audio feature)
    - HMM: Model the emission and transition probabilities
        - Emission : Probability of audio features given phoneme
        - Transition: Probability of transitioning from phoneme at t-1 to phoneme at t
        - Viterbi Decoder to maximize the likelihood of observations, argmax is the best prediction of phonemes. Phonemes → Words later.
    - GMM are easy to implement, can model a lot of data with enough components, but not necessarily to best way to build an acoustic model of the data. Does not work well with nonlinear manifold

What do we need?

“WHAT WE NEED IS A BETTER METHOD OF USING THE INFORMATION IN THE TRAINING SET TO BUILD MULTIPLE LAYERS OF NONLINEAR FEATURE DETECTORS. “ - Hinton et. al

- Stack RBMs → Deep Belief Nets
- Use DBN to replace GMMs in acoustic Modelling so it can represent nonlinear Manifolds better
- Problem (Back then):
    - GMMs were easy to parallelize
    - Training DBN with clusters of machines is not easy as much then.
    - Compute limits



[TABLE 3] A COMPARISON OF THE PERCENTAGE WERs USING DNN-HMMs AND GMM-HMMs ON FIVE DIFFERENT LARGE VOCABULARY TASKS.

TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

Listen, Attend and Spell: A Neural Network For Large Vocabulary Conversational Speech Recognition

W. Chan, N. Jaitly, Q. Le, O. Vinyals

Listen, Attend and Spell (LAS) (2016)

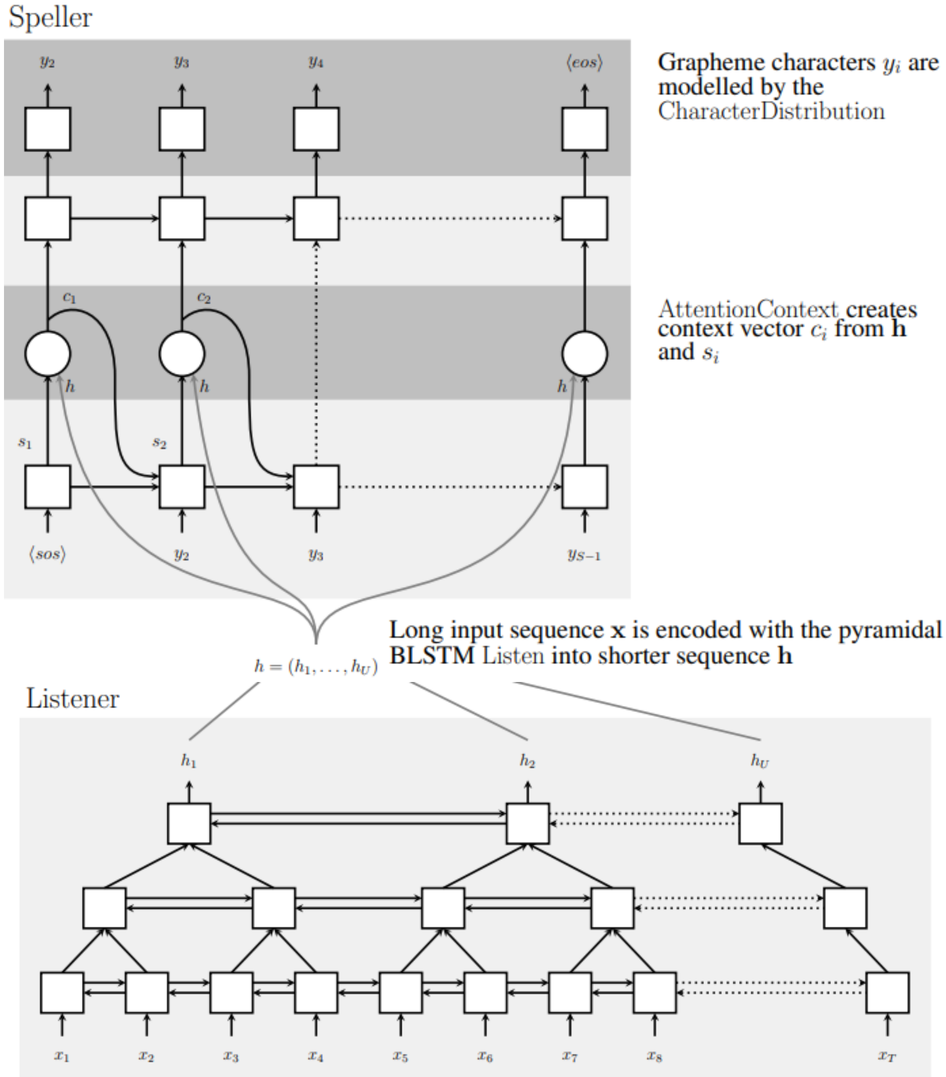
- DNN-HMM → Seq-2-Seq
- Acoustic, Pronunciation & Language models → Single NN jointly learns them
    - Acoustic Models: Emission Probabilities in HMM
        - P(X | h) : Dist. of audio features (observations X) given phones (latent variables h)
        - Phone-to-Acoustic Feature modelling
    - Pronunciation Models: Statistical mapping from words to phones
    - Language models: $P(word_{t+1}|word_t) = P(word_{t+1},word_t)/P(word_t)$
    - LanguageModel → Next word → PronunciationModel → Phones → Acoustic Model → Feats
- Previously trained seperatedly, CTC attempts to go from speech to transcripts end-to-end

Previously on: Joint Representation

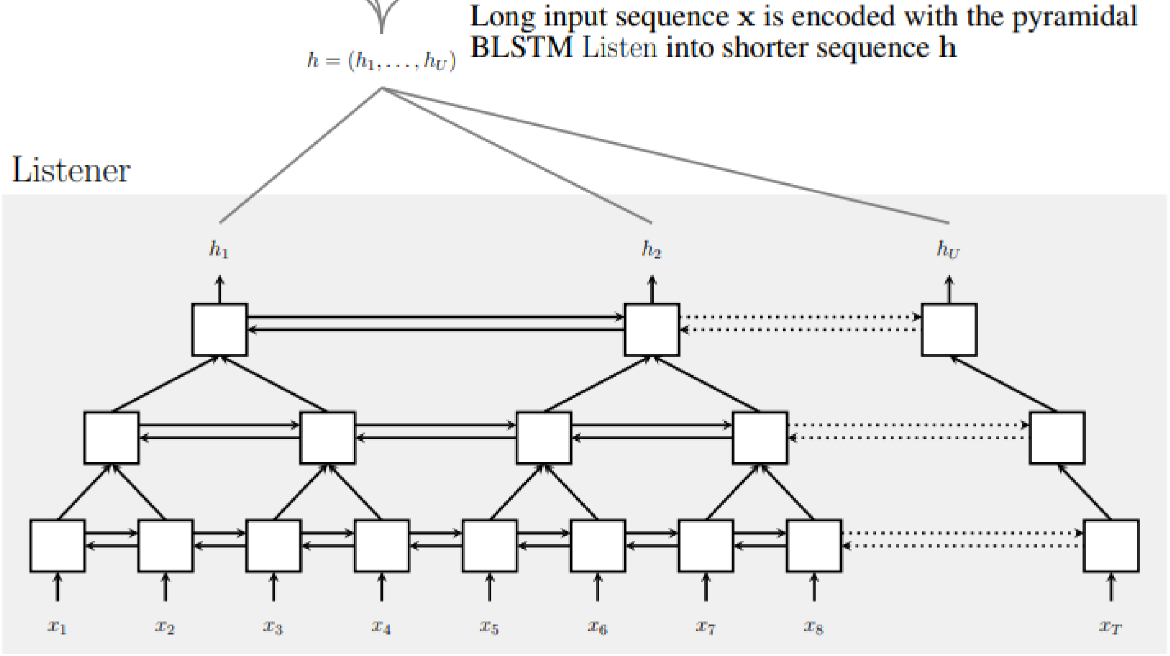
- CTC: Connectionist Temporal Classification
  - Labels are assumed to be conditionally independent of each other.
- Sequence-to-Sequence with Attention
  - Only applied to phoneme sequences
- LAS improves on a number of things:
  - Speller outputs one character at a time: out-of-vocabulary and rare words are handled since the transcription is not word level.
  - LAS can generate multiple spelling variants (e.g. “triple a” → “aaa” + “triple a”)

Listen, Attend, Spell (LAS)

- Context
  - Seq-2-seq models were Encoder-Decoder schemes.
  - Encoder compresses/summarizes final vector in seq. into latent space called Context C
  - Decoder generates the seq. Output from Context C.
- Attention
  - Briefly, instead of just using the last past information, you learn a mask to cherry-pick a subset of the past states
  - You pay more “attention” to some past subsequence in performing prediction



Listen



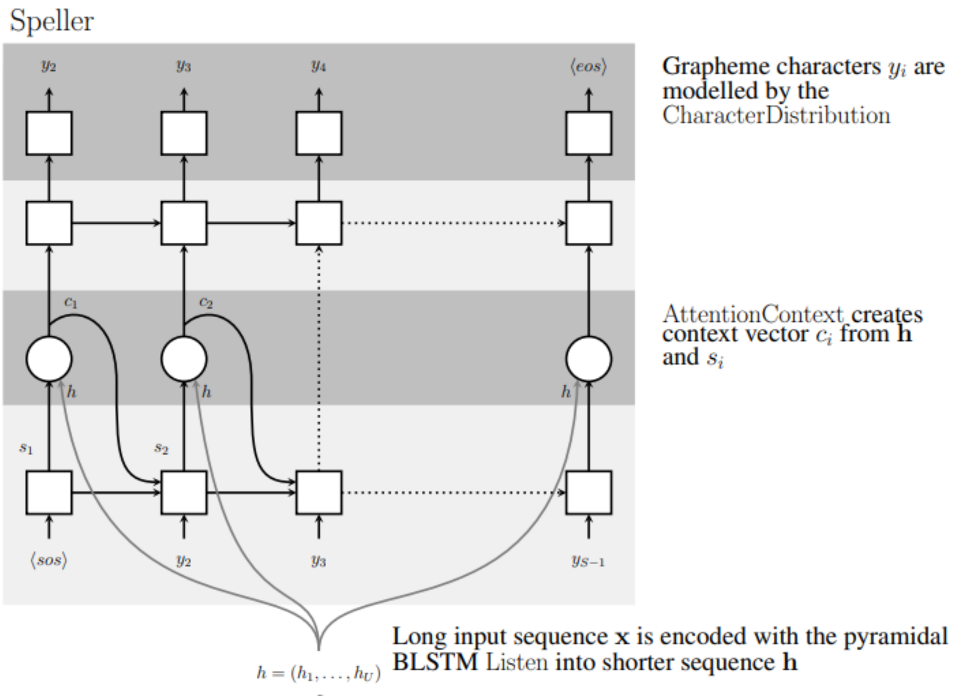
- Pyramid of Bidirectional LSTM
  - Faster convergence, otherwise training takes long.
- Encode input sequence into higher-level features
- Learns the Acoustic Model Encoding

$$\mathbf{h} = \text{Listen}(\mathbf{x})$$

Attend & Spell

- Attention: From listeners’ features and past state, create current context.
    - Context: Acoustic info needed to generate the next character
- Speller: Compute a distribution over characters given context at i

$$c_i = \text{AttentionContext}(s_i, \mathbf{h})$$
$$s_i = \text{RNN}(s_{i-1}, y_{i-1}, c_{i-1})$$
$$P(y_i|\mathbf{x}, y_{<i}) = \text{CharacterDistribution}(s_i, c_i)$$



$$\mathbf{h} = \text{Listen}(\mathbf{x})$$
$$P(\mathbf{y}|\mathbf{x}) = \text{AttendAndSpell}(\mathbf{h}, \mathbf{y})$$

Accent Classification

Motivations for Accent Classification

- Identification of country-of-origin for non-native speakers
- Regional-dialect classification
- First step in accent modification
  - Converting accent of speech audio for ease of comprehension
- Multi-Accented Speech Recognition

Components of Accent

- Spectral characteristics of individual phonemes (articulation)
  - Addition
  - Distortion
  - Omission
  - Substitution
- Prosodic elements
  - Rhythm
  - Intonation and Pitch contour

[1]

Main Approaches

- Hidden Markov Models
  - L. Arslan and J. Hansen 1997
- Unsupervised Learning
  - M. Najafian, A. DeMarco, S. Cox and M. Russell 2014
- Recurrent Neural Networks
  - Y. Jiao, M. Tu, V. Berisha and J. Liss 2016
  - K. Rao and H. Sak 2017

Language Accent Classification in American English

Levent M. Arslan and John H.L. Hansen

HMM Approach

- Set of acoustic features extracted from sampled audio waveform
  - Mel-cepstrum coefficients, energy and first order differences
- Three scenarios tested
  - Isolated Word - Full Search
    - 20 word vocabulary
    - HMM recognizer for each word for each accent
  - Continuous Speech - Full Search
    - Vocabulary Independent
    - Monophone HMM recognizers for each accent type
    - Viterbi decoder determines probability of accent
  - Continuous Speech - Partial Search
    - Same monophone model
    - Testing utterance known a priori

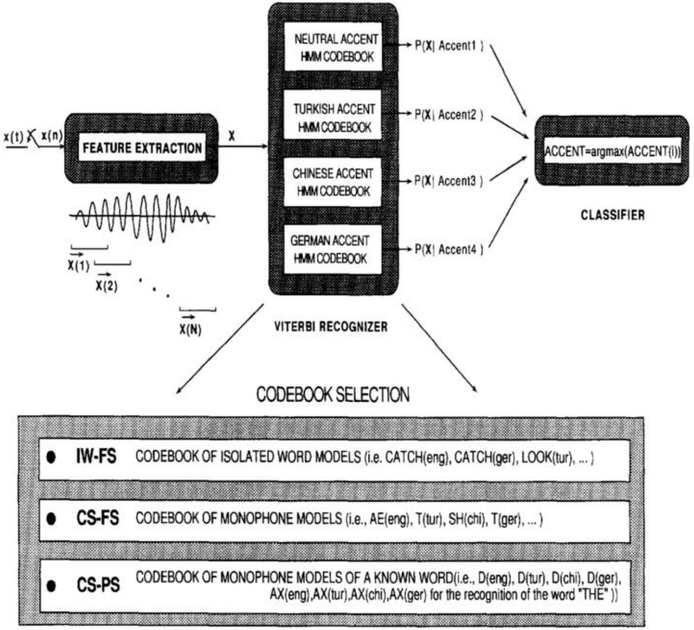


Fig. 5. Framework for the accent classification algorithm.

Evaluation

- IW-FS showed best performance
- Certain words show higher classification rates
- The more words spoken the better the classification
  - One of the main weaknesses of HMM methods

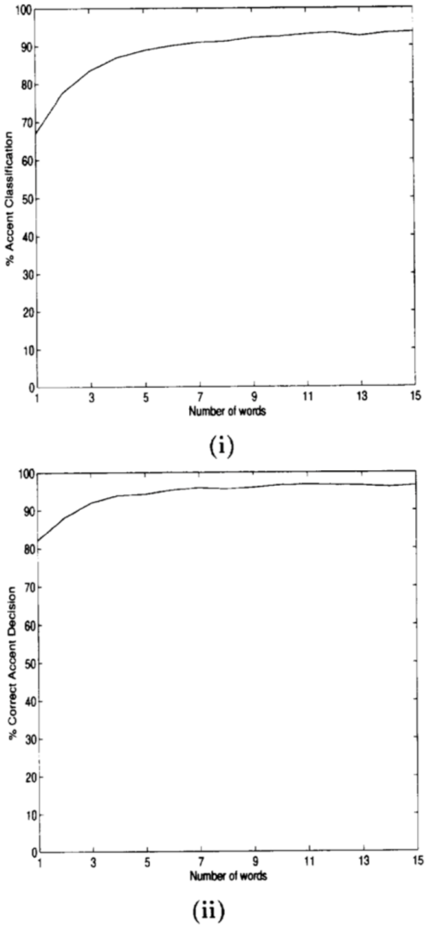


Fig. 7. The effect of speech duration on (i) accent classification and (ii) accent detection rates for 12 open-test speakers among 4 different language accents.

Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features

Yishan Jiao, Ming Tu, Visar Berisha, Julie Liss

DNN and RNN Fusion Approach

- Voice Activity Detection removes silence
- Speech segments split between long term and short term features
  - Short term: 39th order mel-scale filterbank features with logarithmic compression
  - Long term: Mean, standard deviation, kurtosis of MFCC, RASTA
- Deep Neural Network makes prediction from long term features
- Recurrent Neural Network makes prediction from short term features
- Combine DNN and RNN probabilities as a weighted average

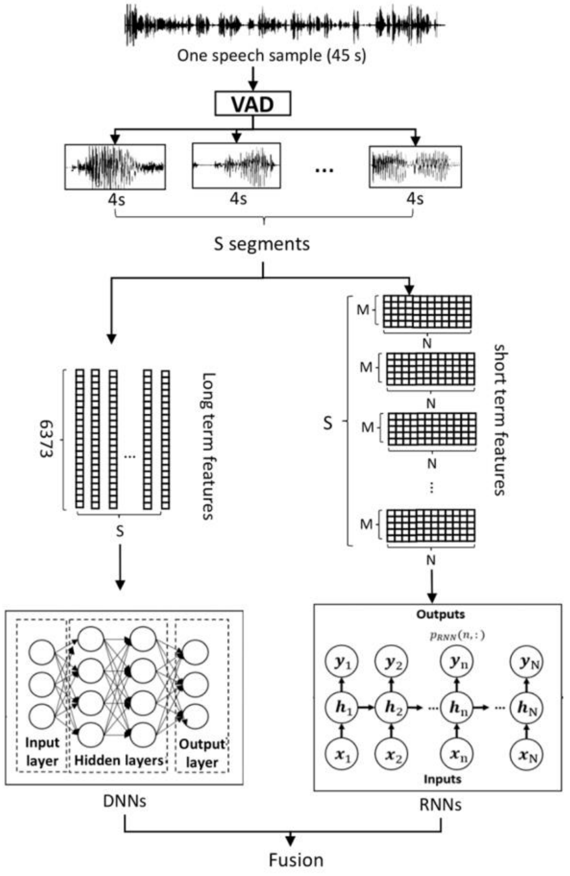


Figure 1: The proposed system of combining long and short term features using DNNs and RNNs.

Evaluation

- 45 second speech samples from 11 native languages
- Trained on 3300 speech samples
- Developed on 965 samples
- Tested on 867 samples
- Best performance when fusing DNN, RNN and baseline SVM-based system
  - Found confusion between accents of people living in geographically close regions

Table 4: Confusion matrix of the proposed system fused with baseline on development set. Rows are reference, and columns are hypothesis.

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	**36**	3	3	8	5	6	3	0	4	6	11
CHI	1	**55**	3	3	4	4	1	7	2	2	2
FRE	9	1	**40**	3	2	9	1	2	8	1	4
GER	3	7	4	**58**	1	5	0	0	1	1	5
HIN	0	1	0	0	**64**	1	2	0	0	14	1
ITA	7	1	5	3	4	**64**	1	0	5	0	4
JPN	3	15	2	0	2	4	**38**	12	8	1	0
KOR	2	21	1	2	2	2	9	**43**	4	1	3
SPA	6	8	8	2	5	9	7	8	**35**	3	9
TEL	2	1	0	2	34	1	0	0	2	**41**	0
TUR	9	2	0	5	3	6	2	1	3	1	**62**

References

1. L. Levent and J. Hansen: “Language accent classification in American English,” *Speech Communication*, Vol. 18, pp 353-367, 1996.

2. Y. Jiao, M. Tu, V. Berisha, and J. Liss: “Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features,” *Interspeech*, pp. 2388-2392, 2016.

3. M. Najafian, A. DeMarco, S. Cox, and M. Russell: “Unsupervised Model Selection for Recognition of Regional Accented Speech,” *Interspeech*, pp. 2967-2971, 2014.

4. Baker, J. (1975). Stochastic modeling for automatic speech recognition, in D. R. Reddy, (ed.), *Speech Recognition*, Academic Press, New York.

5. Huang, X. D., A. Acero, and H. Hon (2001). *Spoken Language Processing—A Guide to Theory, Algorithms, and System Development*, Prentice Hall, Upper Saddle River, NJ.
6. Baum, L. (1972). An inequality and associated maximization technique occurring in statistical estimation for probabilistic functions of a Markov process, *Inequalities*, III, 1–8.
7. Baker, J. (1975). Stochastic modeling for automatic speech recognition, in D. R. Reddy, (ed.), *Speech Recognition*, Academic Press, New York
8. Rao, K., & Sak, H. (2017, March). Multi-accent speech recognition with hierarchical grapheme based models. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4815-4819). IEEE.
9. W. Chan, N. Jaitly, Q. Le, O. Vinyals: “Listen, Attend and Spell: A Neural Network For Large Vocabulary Conversational Speech Recognition,” *ICASSP*, pp. 4960-4964, 2016.
10. G. Hinton et. al. “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *IEEE Signal Processing Magazine*, pp. 2-17, 2012.

Questions?

Thank You