Monophone-based Background Modeling for two-stage on-device Wake Word Detection

Minhua Wu, Sankaran Panchapagesan, Ming Sun, Jiacheng Gu, Ryan Thomas, Shiv Naga Prasad Vitaladevuni, Bjorn Hoffmeister, Arindam Mandal
wuminhua@amazon.com, panchi@google.com, mingsun@amazon.com

Motivation

Goal: To improve the accuracy of the wake word detector on the Amazon device

Focus of this work: Incorporate monophone-based units to model the non-keyword background

Baseline Two-stage Wake Word System

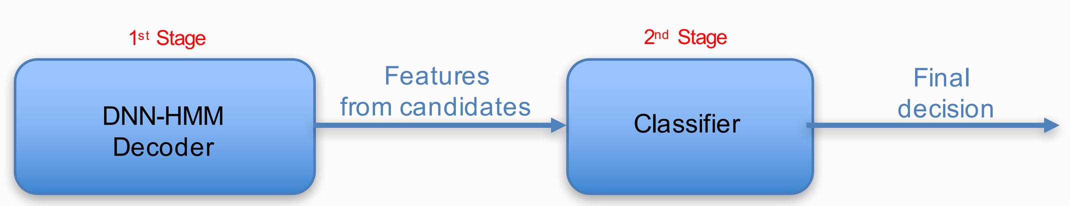

Figure 1: The two-stage wake word detector

1st Stage DNN-HMM Decoder

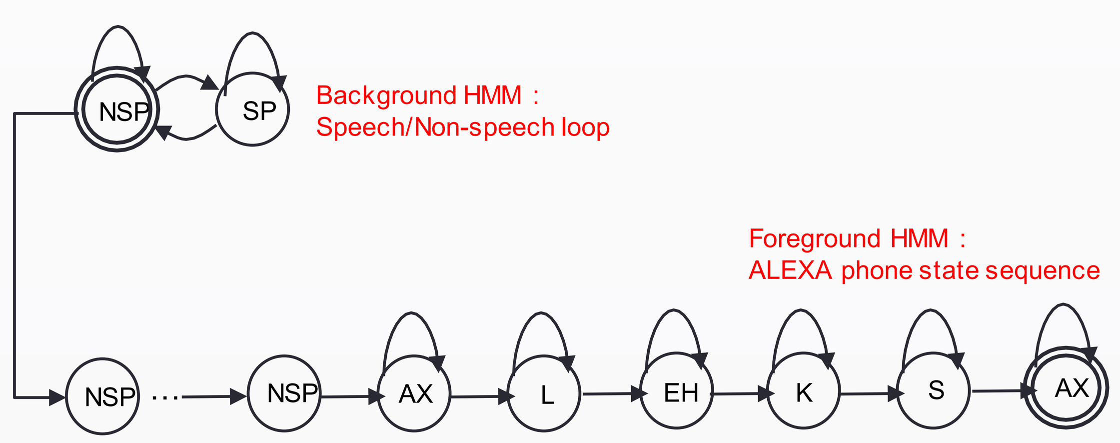

Figure 2: A simplified 1st stage HMM decoding graph for the wake word "Alexa"

- Foreground HMM: wake word phone states
- Background HMM: speech and non-speech states loop
- Acoustic Model: Deep Neural Network (DNN)
- Decoder: Viterbi decoding on the graph
- Wake Word Hypothesized: When difference in foreground and background log likelihoods exceeds a threshold
- 1st stage DET curve: Tune weight on arcs and states

2nd Stage Classifier

- Second Stage Feature Vector
  - Obtained from 1st stage wake word hypothesis
  - Captures info from the whole candidate segment (e.g. segment duration, likelihood etc.)
  - Captures info related to each phone segment

    (e.g. phone duration, confidence scores etc.)
- Use a small feed-forward Neural Network (NN) for experiments

New Wake Word System Using Monophone-based Background Modeling

New 1st Stage DNN-HMM Decoder

- New Background HMM:
  - Expand speech, non-speech events to various monophones
  - Becomes a phone-level unigram FST
- New Acoustic Model: background targets expanded

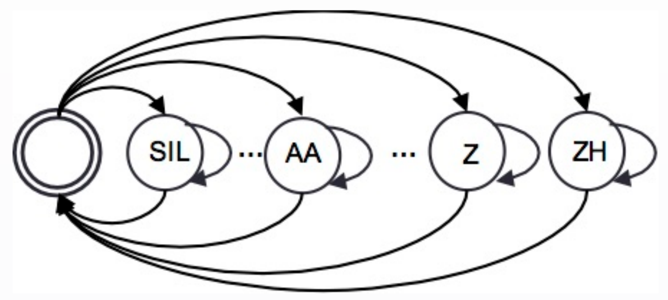

Figure 3: A simplified 1st stage monophone-based background HMM. 3-state HMM topology is actually used

New Feature Engineering for 2nd Stage Classifier

- Baseline second stage features are still valid
- Extra Features: new scores measuring the degree of match between each candidate's wake word phone segment p and every background monophone q.

$p \in$ wake word phones: $\{SIL_{Preceding}, AX\_B_{Alexa}, L_{Alexa}, EH_{Alexa}, K_{Alexa}, S_{Alexa}, AX\_E_{Alexa}\}$
$q \in$ background monophones: $\{SIL, SPN, NSN, PAU, AA, AE, ...Y, ZH, Z\}$

$$MatchScore_{p,q} = \frac{1}{Dur_p} \sum_{t=T_p}^{T_{p+1}-1} max\{log(P(X_t|Q_q^L, \theta_{BG})), log(P(X_t|Q_q^C, \theta_{BG})), log(P(X_t|Q_q^R, \theta_{BG}))\}$$

  - p: A wake word phone (T_p: start frame of this phone in the candidate segment X)
  - Q_q^L, Q_q^C, Q_q^R: The three states for each monophone q.
  - Obtain match score for each candidate's wake word phone p with respect to every background monophone q
  - Distinguish better between real wake words and confusable segments among first stage candidates

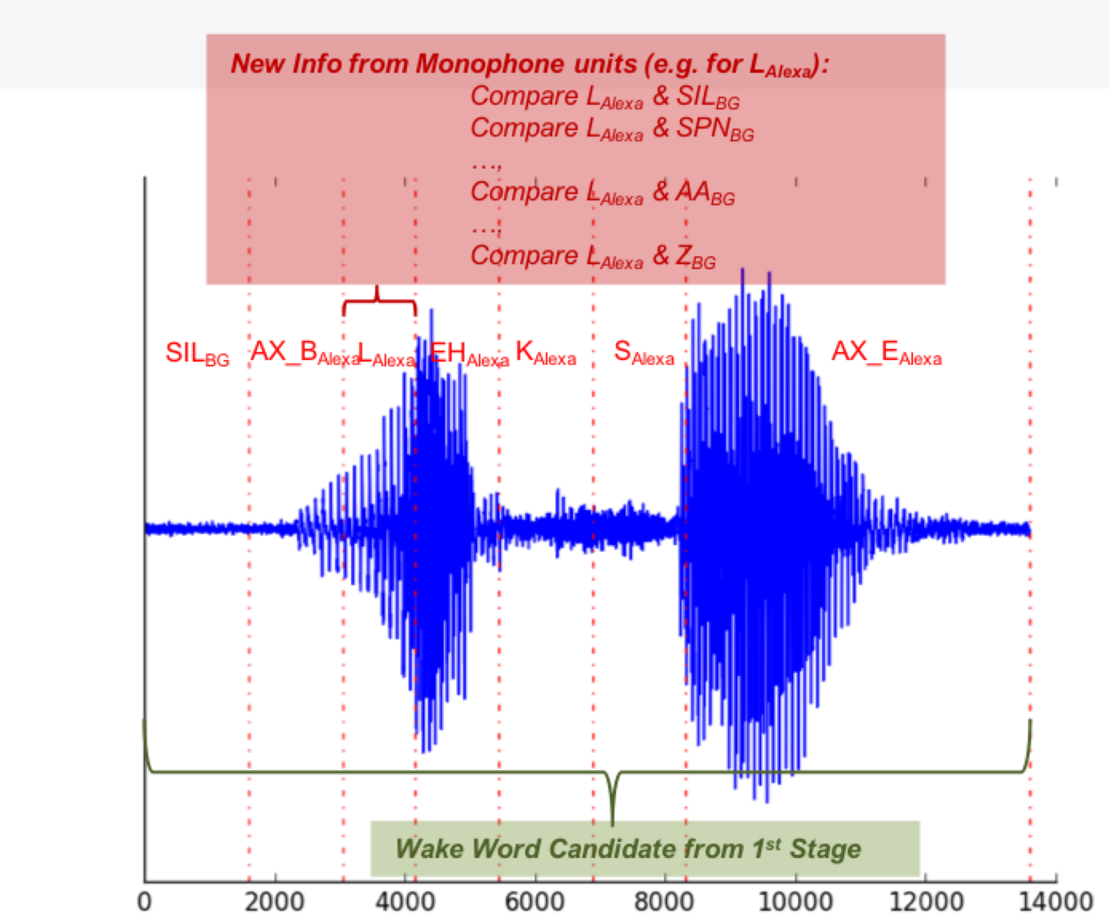

Figure 4: Extracting extra information from first stage wake word hypothesis using monophone based units for background

Experiment Results

Baseline Setup

- Several thousand hours of real far-field data for training
- Approximately 30,000 wake word instances in dev/test set
- A feed-forward DNN acoustic model at the first stage
- Features: Log Mel-Filter-Bank Energies (LFBE) (20 frames for left context and 10 frames for right context)
- Wake word task (50 targets) multi-task trained with LVCSR targets [1]
- The GPU-based distributed DNN trainer utilized [2]
- The second stage feature vector is of dimension 67
- A small feedforward NN as the second stage classifier

Changes for Monophone-based System Setup

- Use 44 monophones in the background model
- Background HMM changes to be a phone-level unigram FST
- DNN output targets for the wake word task is expanded
- The second stage feature vector is of dimension 375 (67+7x44)

1st Stage HMM Tuning

- Performance: The two systems are almost the same at this stage
- Operating points picked for building 2nd stage classifier:
  - recall at around 0.02 for both systems

End-to-end Evaluation

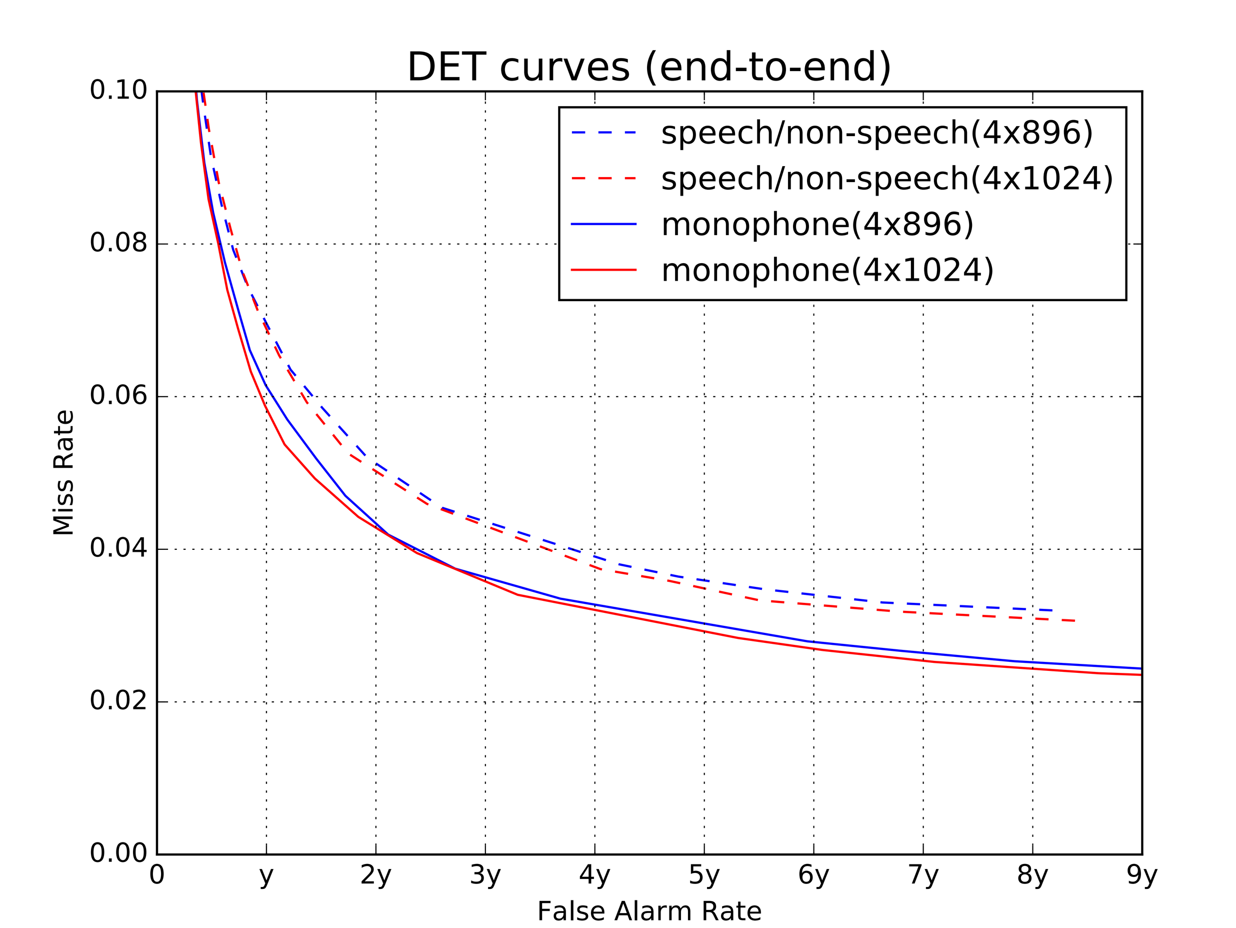

Figure 5: End-to-end comparison of the baseline wake word system and the new system using monophone-based background modeling (1st stage DNN size recorded in the legend, 2x64 for 2nd stage NN); DET curves on test set; Axis of the false alarm rate is obscured due to the sensitive nature of this information

Table 1: Summary of different wake word systesms

	2nd NN: 2x64		
	FRR (Fix FAR=2y)	FAR (Fix FRR=0.04)	# of Params
SP/NSP(4×896)	0.051	3.71y	3.02M
SP/NSP(4×1024)	0.050	3.43y	3.84M
Monophone (4x896)	0.043	2.35y	3.15M
Monophone (4x1024)	0.042	2.31y	3.99M

Effectiveness of the 2nd stage

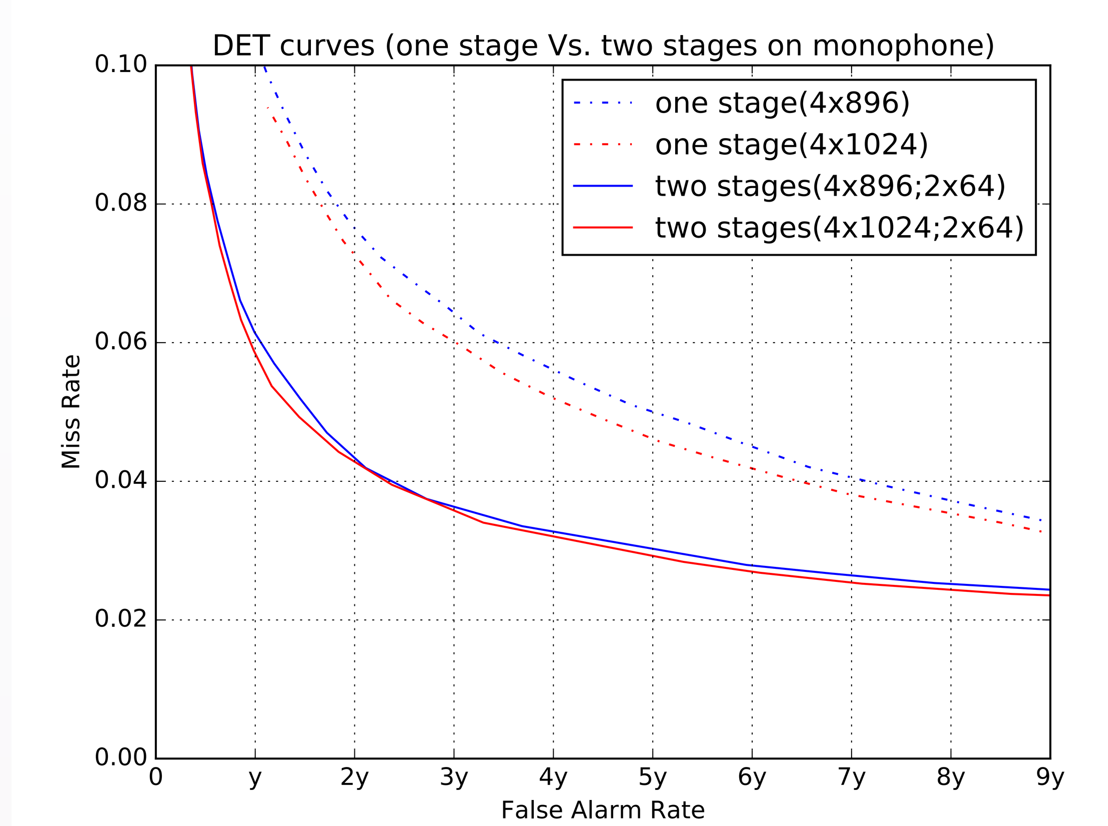

Figure 6: Comparison of the performance with and without 2nd stage classifier (2×64 NN); DET curves on test set; Axis of the false alarm rate is obscured due to the sensitive nature of this information

Conclusion

- Propose a new way to model the non-keyword part.
  - Expand the speech/non-speech events to more specific monophone-based units at the first stage.
  - Extract extra match scores for final detection
- The new system reduces FAR by **37%** when the FRR level is maintained.
- On the other hand, it reduces FRR by about **16%** if FAR level is fixed.
- The second stage itself is able to reduce FAR by **67%** relatively on top of 1st stage hypothesis.

References

[1] Sankaran Panchapagesan, Ming Sun, Aparna Khare, Spyros Matsoukas, Arindam Mandal, Bjö rn Hoffmeister, and Shiv Vitaladevuni. Multi-task learning and weighted cross-entropy for dnn-based key- word spotting. *Interspeech 2016*, pages 760–764, 2016.

[2] Nikko Strom. Scalable distributed dnn training using commodity gpu cloud computing. In *INTER- SPEECH*, volume 7, page 10, 2015.