Speech-In-The-Wild Analytics in the Era of Deep Learning:

Recent Advancements and Remaining Challenges

Dimitra Vergyri

SRI International

With contributions from:

SRI: Mitchell McLaren, Horacio Franco, Martin Graciarena, Aaron Lawson, Diego Castan, Mahesh Nandwana, Julien Van Hout, Colleen Richey

UBA-CONICET: Luciana Ferrer

Apple/UMD: Vikramjit Mitra

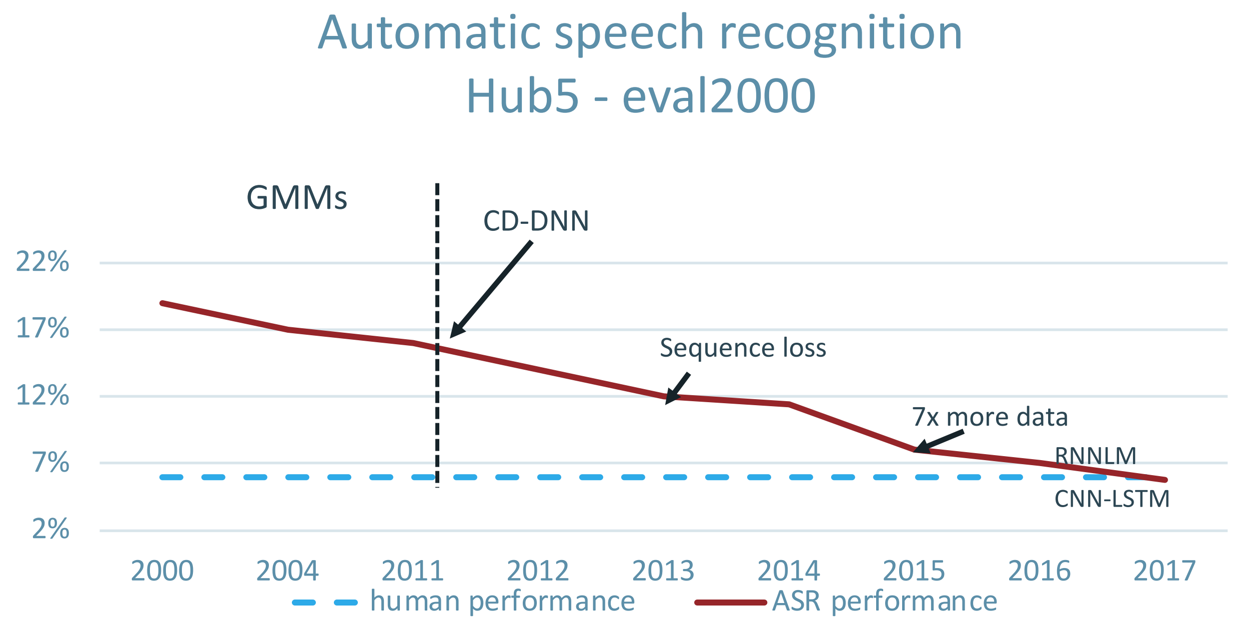
2018 IEEE Spoken Language Technology Workshop

Impact of Deep Learning on Speech Processing

Deep learning solved speech processing

Huge improvements – especially given

- Well-defined task and conditions
- Large matched-condition datasets
- Focused community effort over long period of time



Problem not completely solved …

Challenges still remain

Important Challenges for Speech Analytics

From controlled recording conditions:

To Audio in the Wild:

Audio in the Wild:

- Signal characteristics
  - Degraded signals
    - Distant microphones, distorted and noisy channels, reverberation, compression, etc.
  - Variability
    - Multiple speakers, speaker states and environments
    - Nonstationary noises and distortions
    - Unexpected events
- Test conditions
  - Mismatch with training
  - Short duration test samples, e.g. 1-5 sec

Addressing Real-world Challenges

Technology needs to work in real-world conditions

- Realistic Datasets
  - Exhibiting a variety of conditions present in real situations
- Research Directions
  - Data augmentation
  - Feature design and learning
  - Deep learning models
  - Adaptation/calibration with limited data
- Example Speech Analytics Tasks
  - Speech Activity Detection (SAD)
  - Speaker ID (SID)
  - Keyword Detection / Query by Example (QbE)

SRI International®

Realistic Datasets

Exhibiting a variety of conditions present in real situations

Datasets Available to the Community

- NIST and other formal evaluations: datasets with challenging conditions
  - CHiME challenges (2011-2018): speech on speech, speech in noise, distant speech
  - NIST SRE eval data (1996-2018): telephone and interview speech data, including non-English
  - DARPA RATS data (2011-2014): mostly PTT, severe transmission noise
- Contributions from the research community
  - ICSI meeting corpus (2004): multiple speakers, close-talking or distant mics
  - VoxCeleb (2017, 2018) : 1000's of celebrity speakers in the wild
- SRI's recent contributions (SITW, VOICES)
  - Many speakers
  - Multiple speaker segments
  - Wide range of "in the wild" artifacts

Speakers In The Wild (SITW) 2016: A "Sample" of Real Conditions

M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," Interspeech 2016

Focus: Multi-speaker, cross-condition data from real-world recordings

Source: Open-source videos

Subjects: 299 public figures

Language: English (native and non-native)

Publicly Available for Research Purposes: www.speech.sri.com/projects/sitw/

Combination of Conditions: e.g.: red carpet, Q&A in auditorium, ice bucket challenge

Targeted conditions:

- Traffic noise
- Reverb
- Multi-layer compression
- Multi-speaker (1-8)
- Conversational speech
- Laughter
- Phone channel
- Background music
- Non-linear effects
- Natural Lombard effect
- Crowd noise
- Restaurant noise
- Variable duration (6 sec – 2 hours)

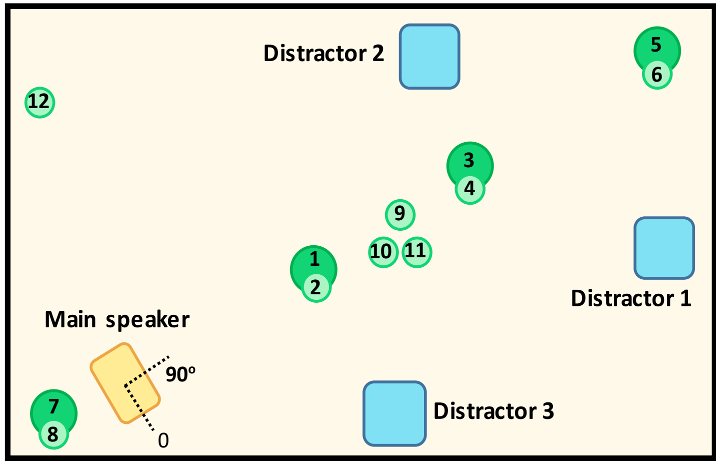
VOICES: voices.lab41.org free download

***Focus:* Distant** microphone recordings – variable but controlled conditions

Source: 300 speakers from LibriSpeech audio (open source), re-recorded in furnished rooms with background noise

C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, M. Graciarena, A. Lawson, M. K. Nandwana et al., "Voices obscured in complex environmental settings (VOICES) corpus," INTERSPEECH 2018

Microphone and loudspeaker placement in one of the collection rooms



- 4 studio mics (larger circles)
- 8 lavalier mics (smaller circles)
- Mic #9 is under a table
- Mics #10 & #11 are attached to the ceiling
- Mic #12 is in the wall
- Speaker can move/rotate

- 4 background noise conditions played from distractor speakers
  - Music and TV noise from single loudspeaker
  - Babble noise from all 3 loudspeakers
  - 15dB SNR measured near mic 1

- 1440 h of retransmitted distant audio (120h/mic)

SRI International

Need Additional Datasets and Robust Learning Approaches

- Deep learning approaches benefit from large amounts of data
- Need **realistic** datasets with wide range of extrinsic and intrinsic variability
- For example:
  - Outdoor collections
  - Longitudinal data of same speakers
  - Intrinsic speaker variability (emotion, voice projection, health, style)
    - E.g. SRI-FRTIV (2009) corpus focused on speaking effort

➢ Need learning approaches that use less data and can generalize to unseen conditions

Research Directions for Robust Speech Analysis

- *Data augmentation*
- *Feature design*
- *Deep learning and feature learning*
- *Adaptation/calibration*

Data Augmentation

- Real-world data often more diverse than development corpora
- Augment corpora to compensate
  - Fabricate data: re-record or simulate channel/reverberation
  - Process signals to simulate variability: vocal tract length variations, speech rate changes, channel effects, etc.
- Successful when target properties are known and can be simulated (e.g. reverberation challenges)
- Hard to generalize to unseen/unexpected conditions

T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in Proceedings of INTERSPEECH, 2015

SRI International

Feature Design for Noise Robustness

V. Mitra, H. Franco, R. M. Stern, J. Van Hout, L. Ferrer, M. Graciarena, W. Wang, D. Vergyri, A. Alwan, and J. H. Hansen, "Robust features in deep-learning-based speech recognition," in New Era for Robust Speech Recognition. Springer, 2017, pp. 187–217

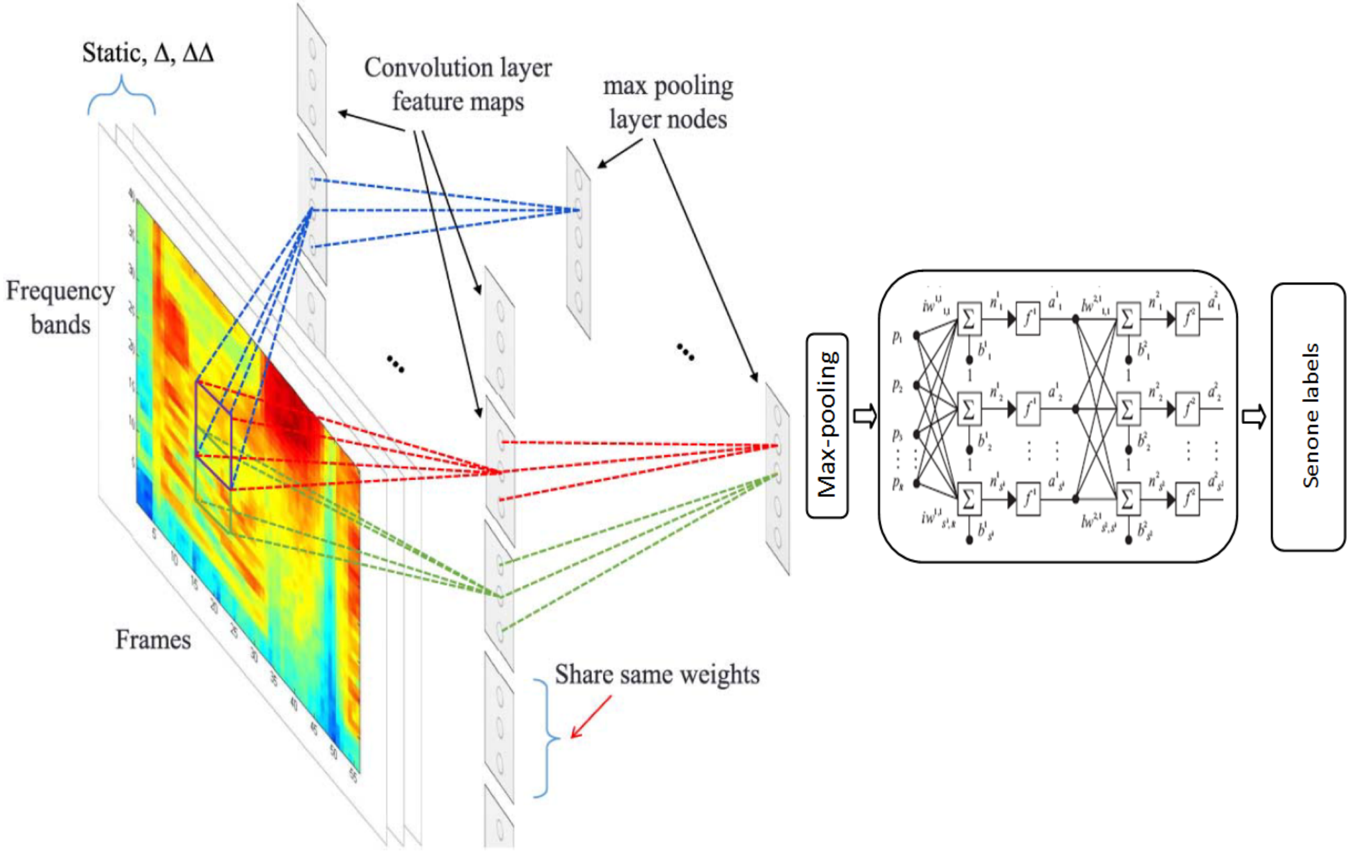
Examples of auditorily and perceptually motivated noise-robust features developed under RATS (2010-2014)

Feature	Site	Characteristics
PNCC	CMU	Uses power-law nonlinearity and noise suppression
Gabor	ICSI	Inspired by high level features discovered in cortical regions of the brain
DOC/SYDOC	SRI	Uses damped oscillator and synchrony processing
NMC/MMeDuSA	SRI	Uses modulation spectrum and root compression
MHEC	UTD	Perceptual MVDR; quantile cepstral dynamics normalization
MbCombF0	UCLA	Variable frame rate analysis; temporal modulation processing; compressive sensing

- Each has advantages
- All designed to work in noisy channels
- Parameters heavily optimized on data
- Combining features often improves results – no single winner

Deep Learning Models Increased Robustness

- Deep Neural Nets (DNNs) significantly improved ASR
- Deep Convolutional Neural Nets (CNNs) emerged as an alternative demonstrating robustness to background noise

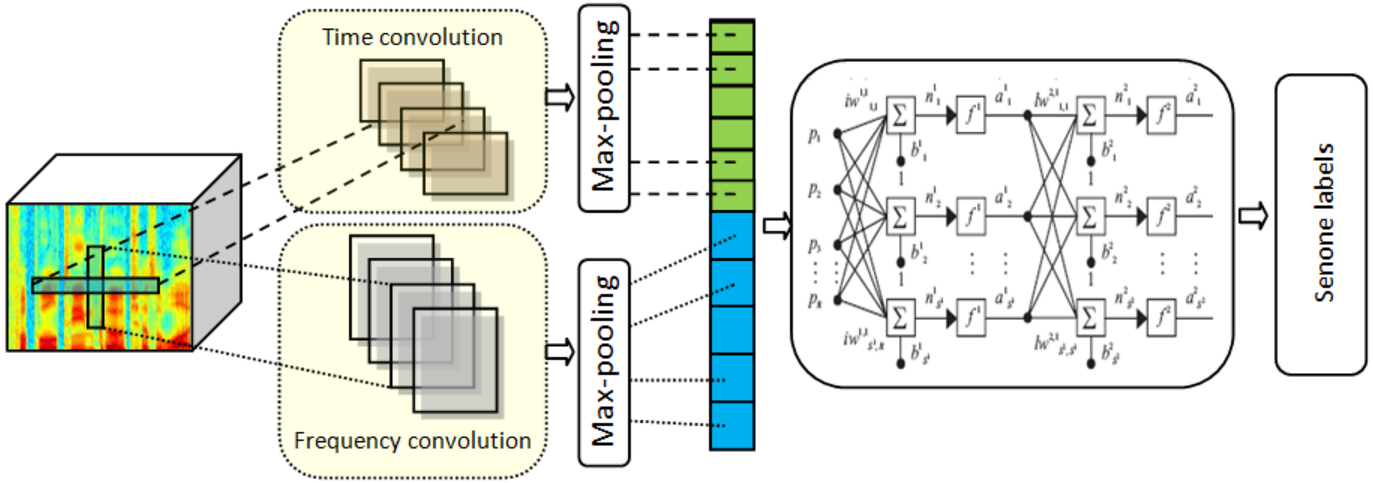


Huang, J., Li, J., & Gong, Y. "An analysis of convolutional neural networks for speech recognition", ICASSP, 2015.
Y. Qian and P. C. Woodland, "Very deep convolutional neural networks for robust speech recognition," in IEEE SLT, 2016

CNN variants: Time-Frequency Convolution (TFCNN)

Several architectures have been explored to improve robustness

- SRI proposed TFCNNs in IARPA ASpIRE challenge
  - Convolution performed across both time and frequency scales
  - Worked very well combined with noise-robust features



V. Mitra and H. Franco, "Time-frequency convolutional networks for robust speech recognition," in Proc. ASRU, 2015.

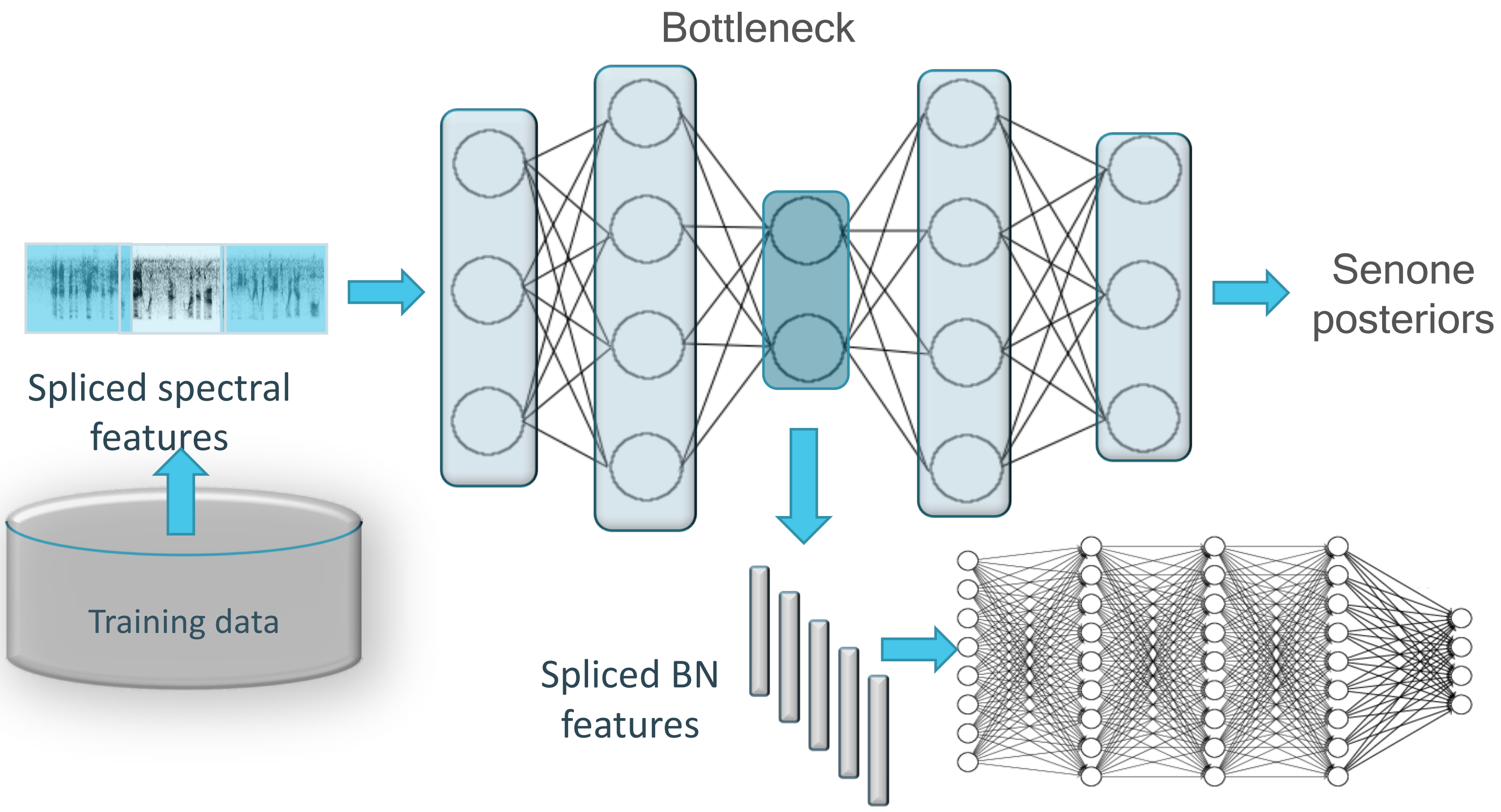
SRI International

Features based on deep-learning models

Bottlenecks (BNs)
- replacing noise-robust features

L. Bai, P. Jančovič, M. Russell, and P. Weber, "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics," in Proc. Interspeech 2015

Bottleneck features discriminatively learned through supervised DNN models

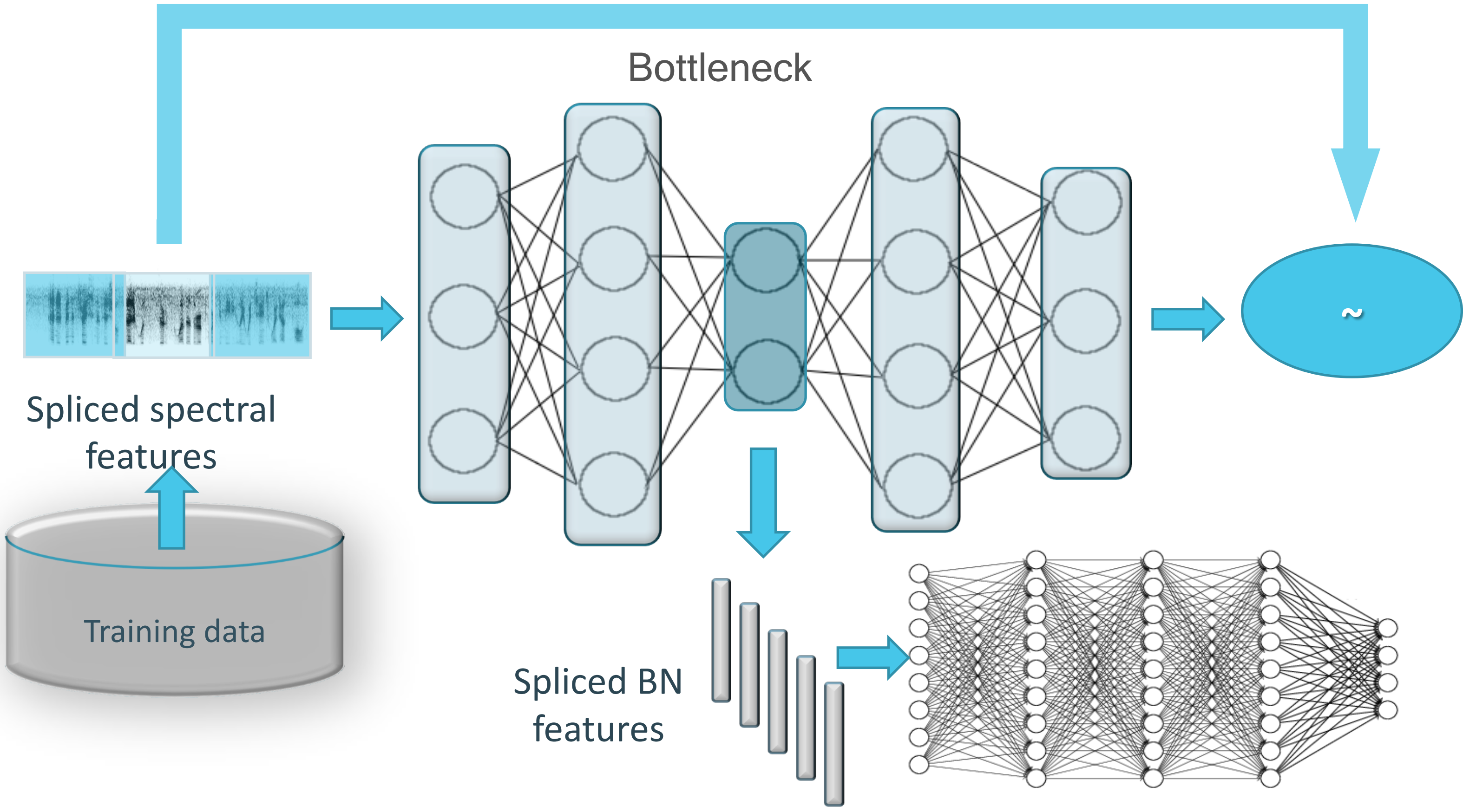


SRI International

Features based on deep-learning models

Bottlenecks from autoencoders

BN features learned through supervised DNN models or unsupervised autoencoders



Stacked BNs have also been used very successfully for multiple tasks

SRI International

Features based on deep learning models

Acoustic embeddings:

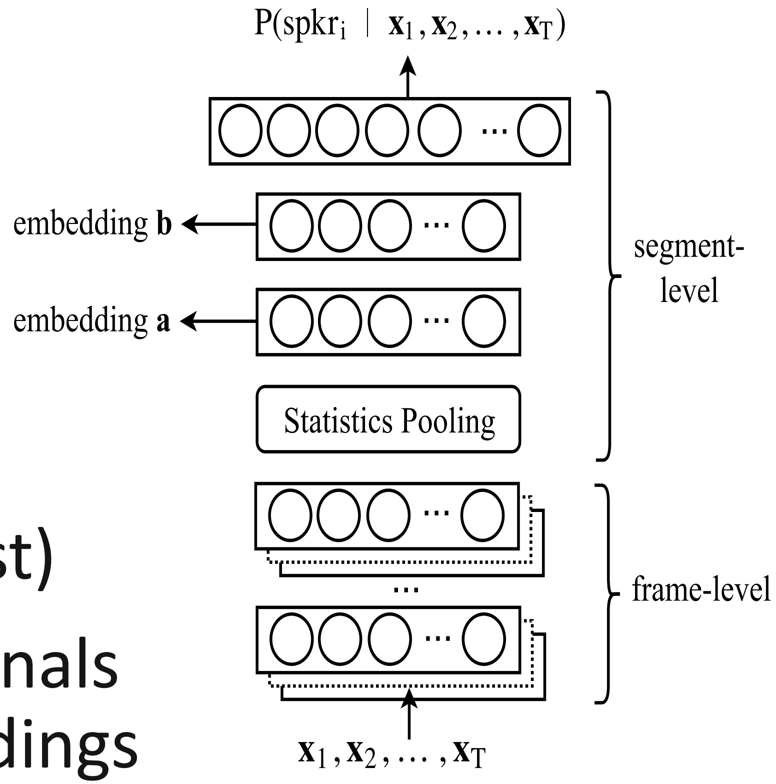
Bottlenecks are frame level, while embeddings learn to map variable-length segments to fixed-length vectors

K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in Proc. ASRU, 2013

M. Rouvier, P.-M. Bousquet, and B. Favre, "Speaker diarization through speaker embeddings," in Signal Processing Conference (EUSIPCO), 2015

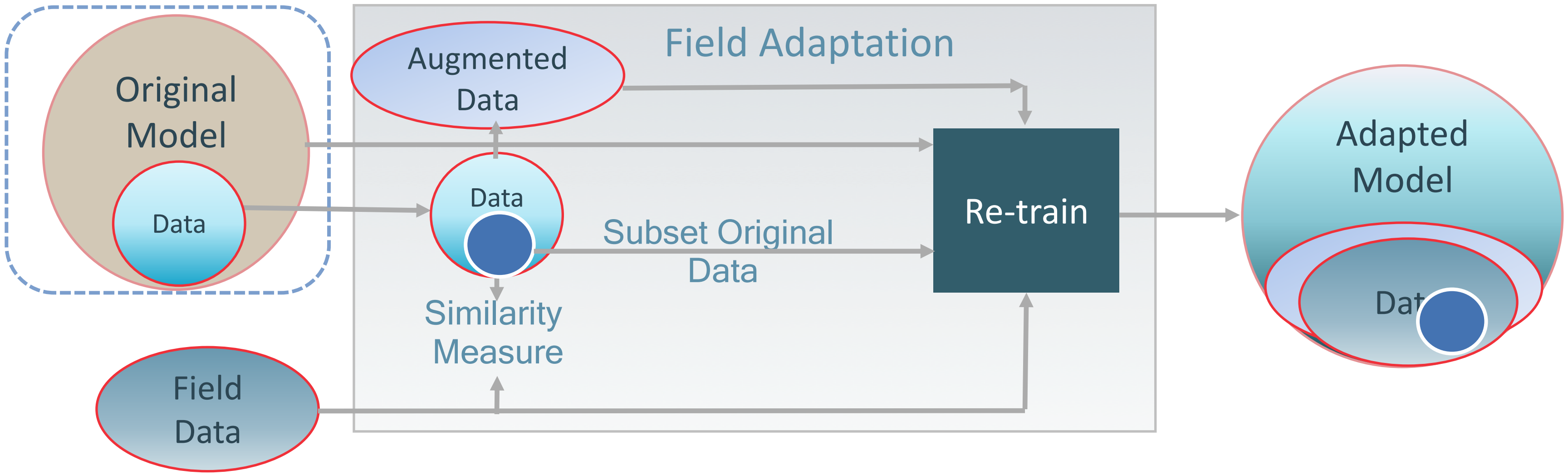
D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition,", ICASSP, 2018

- Embedding layers capture information relevant to a task while removing distortion that may be present in the feature space (robust)
- Similar classes of signals have similar embeddings
- Typical input is MFCCs – have also used PNCCs



Example architecture for speaker ID embeddings (Snyder et al., 2018)

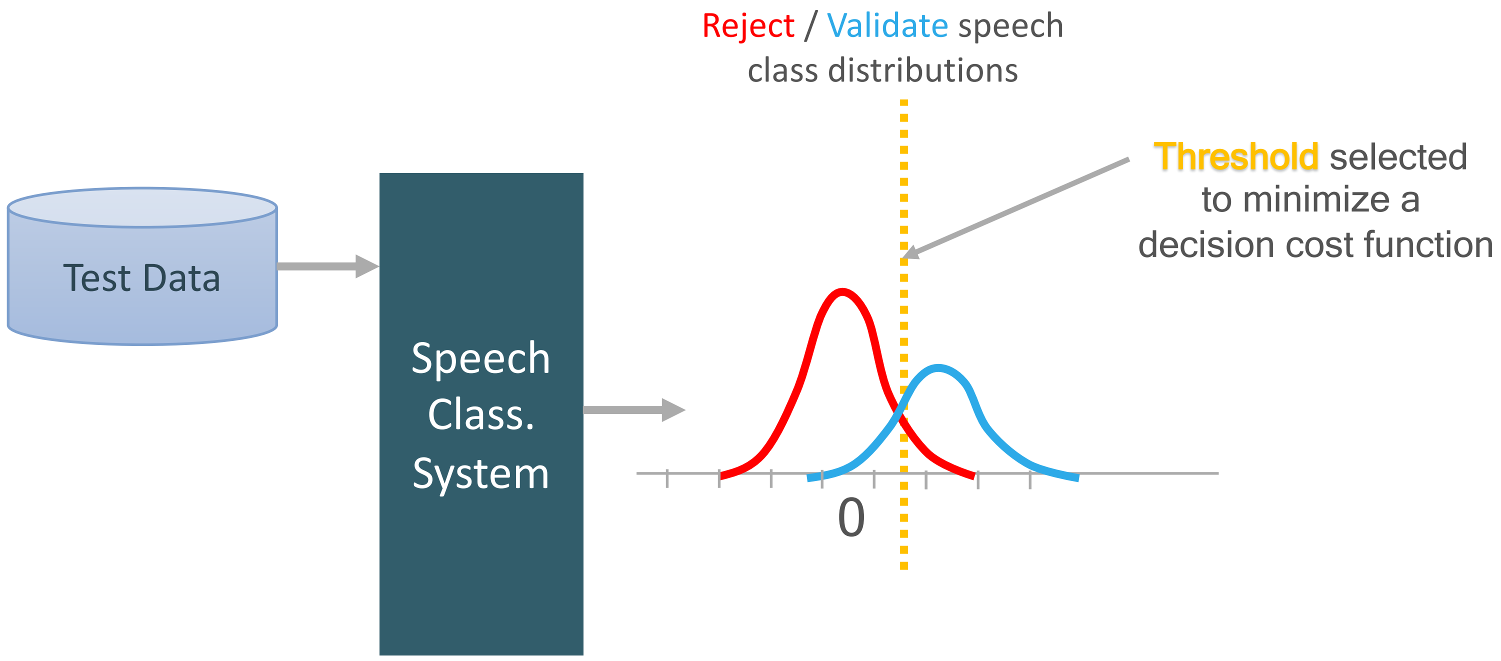
Adaptation to Encountered Conditions



- Exploit available information – including original training data and test data - in the best way possible
  - Unsupervised adaptation: applied when new, unseen conditions are encountered and there are no labels
  - Data selection: dynamically sub-select the most appropriate original data from inside the model
  - Data augmentation: mimic target conditions
- Re-training typically involves only the last layers of the model

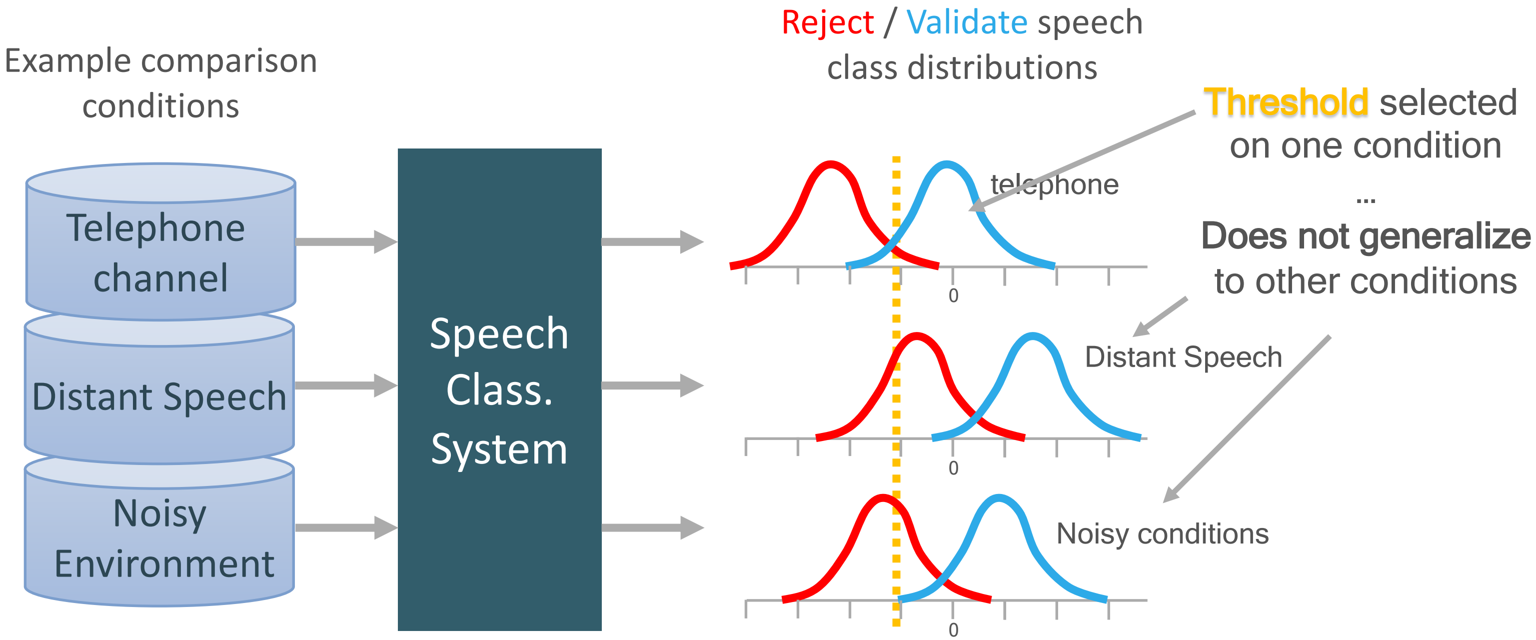
Need for Calibration

Classification decisions are made by applying a threshold to system scores



Need for Calibration

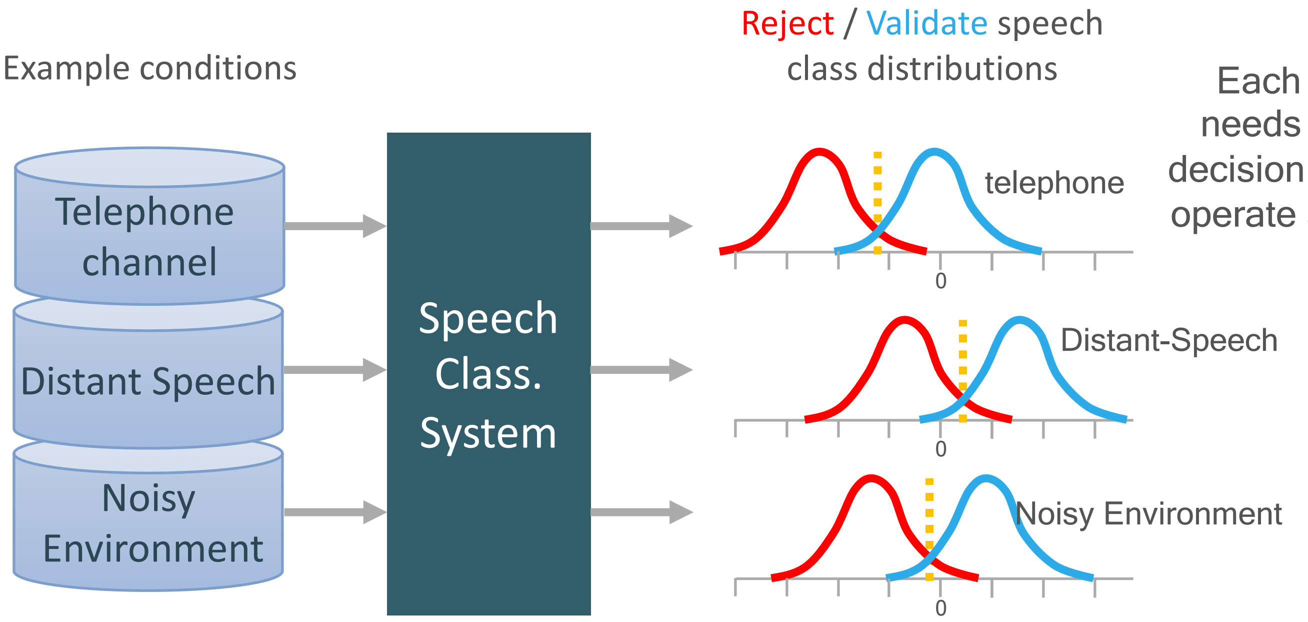
- With signals in the wild, scores **shift** across different conditions
- With a single threshold, system performance on unseen conditions becomes **UNRELIABLE**



SRI International

Need for Calibration

In an ideal world, we would pick a decision threshold for each possible condition…



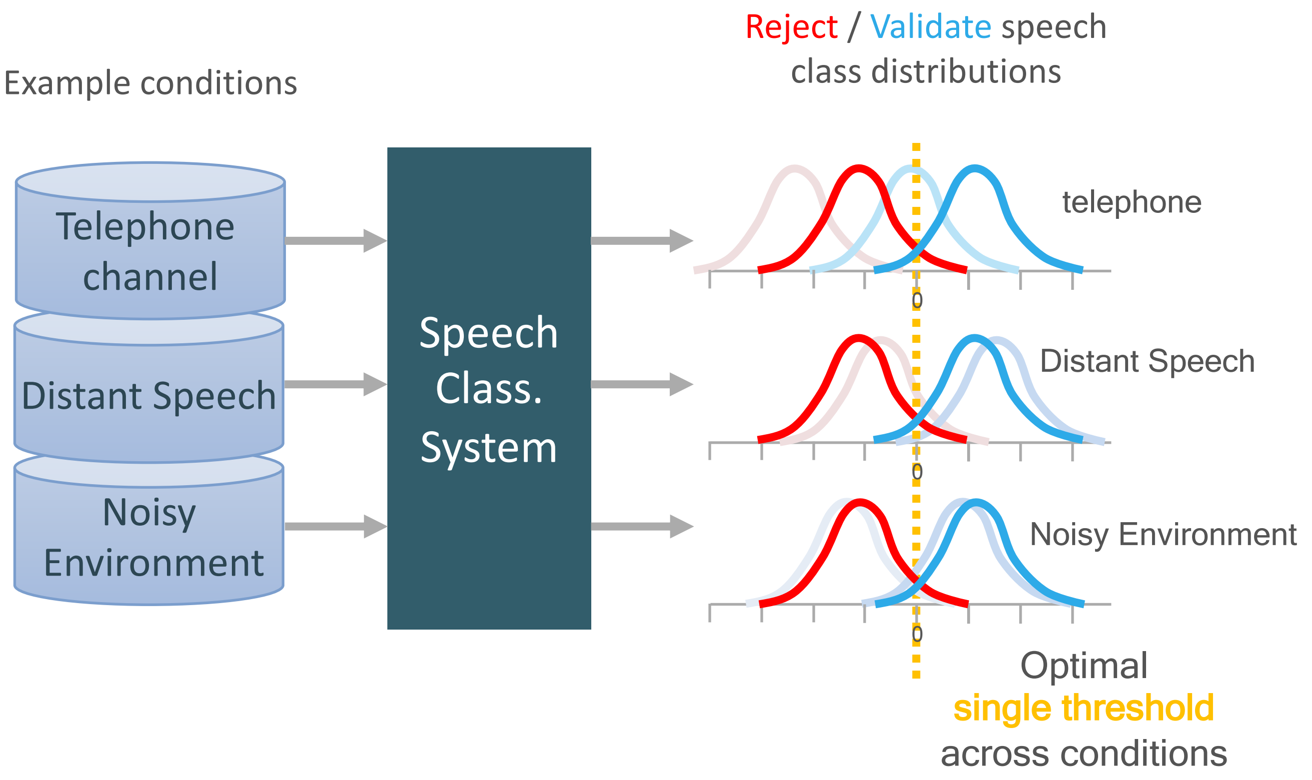
Each condition needs a different decision threshold to operate as expected

Need for Calibration

Must calibrate scores to a common space

- Then a single threshold can be applied with confidence across all conditions

N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection, " Computer Speech & Language 20, no. 2-3 (2006): 230-275



SRI International

Example Speech Analytics Tasks

- *Speech Activity Detection (SAD)*
- *Speaker Identification (SID)*
- *Keyword Detection / Query by Example (QbE)*

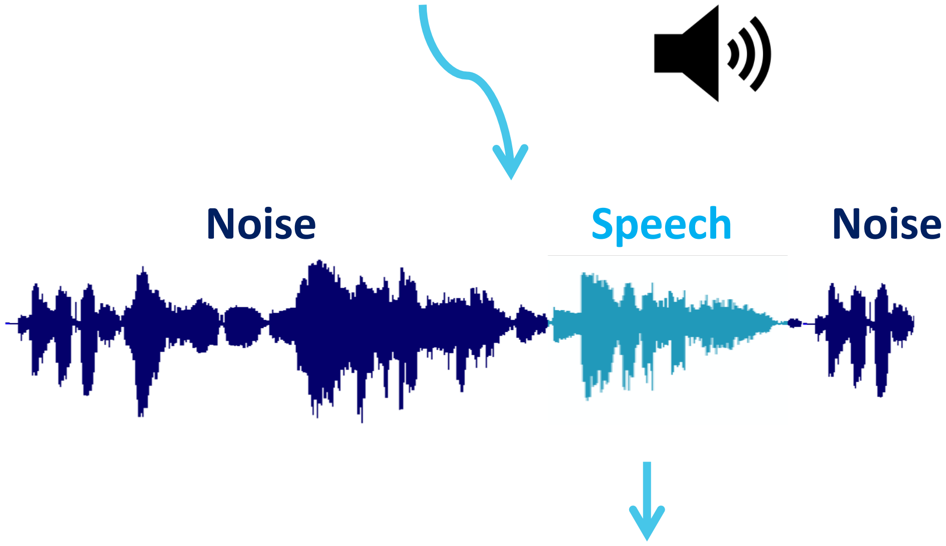
Speech Activity Detection

Goal: detect presence and temporal location of speech in audio signals

- Easy in clean conditions
- Gets harder as environment or channel get noisier

Towards a fast, effective and robust solution

To process speech you first need to find it!



Downstream automated speech processing or human listeners

SAD significantly reduces the amount of data that needs to be processed by a smart interactive device or data mining system

Critical component for follow up systems:

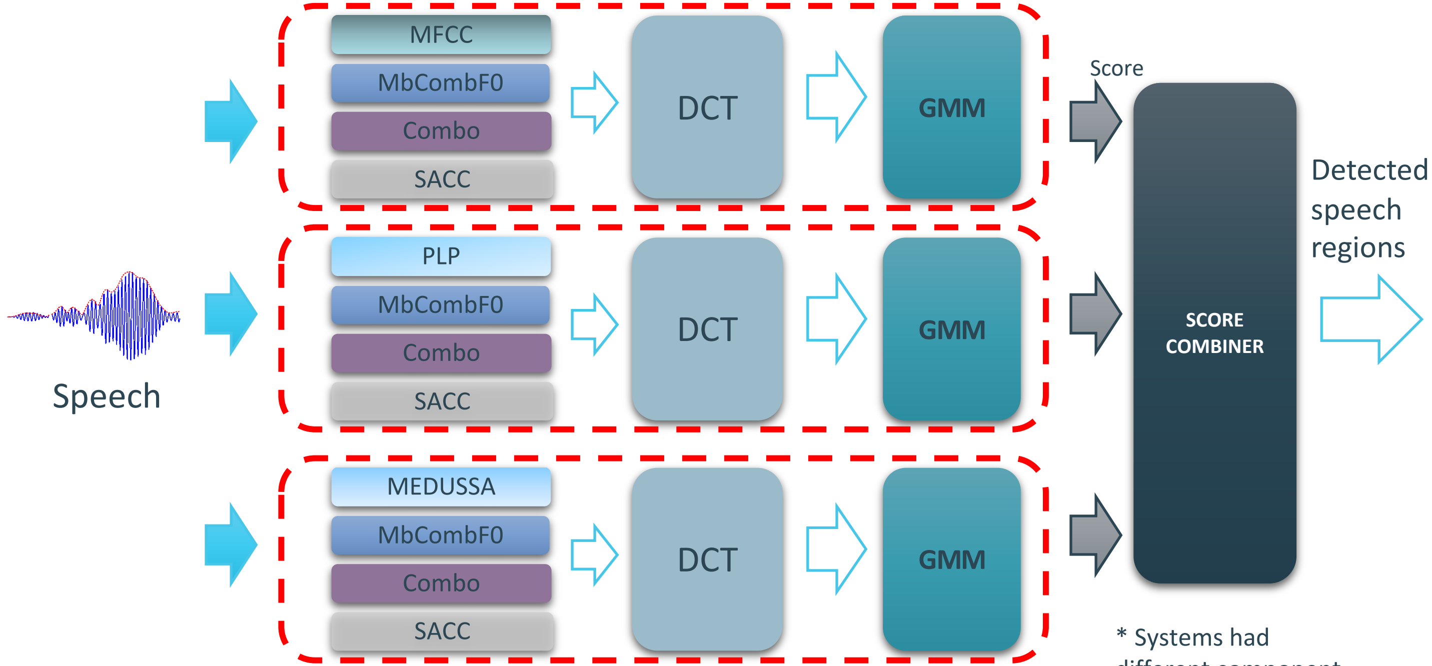
If SAD misses speech segment, no info can be extracted

If SAD false alarms, almost certain error later in pipeline

2014 SRI SAD Pipeline

Noise-robust features + GMMs

Complex multi-system combination

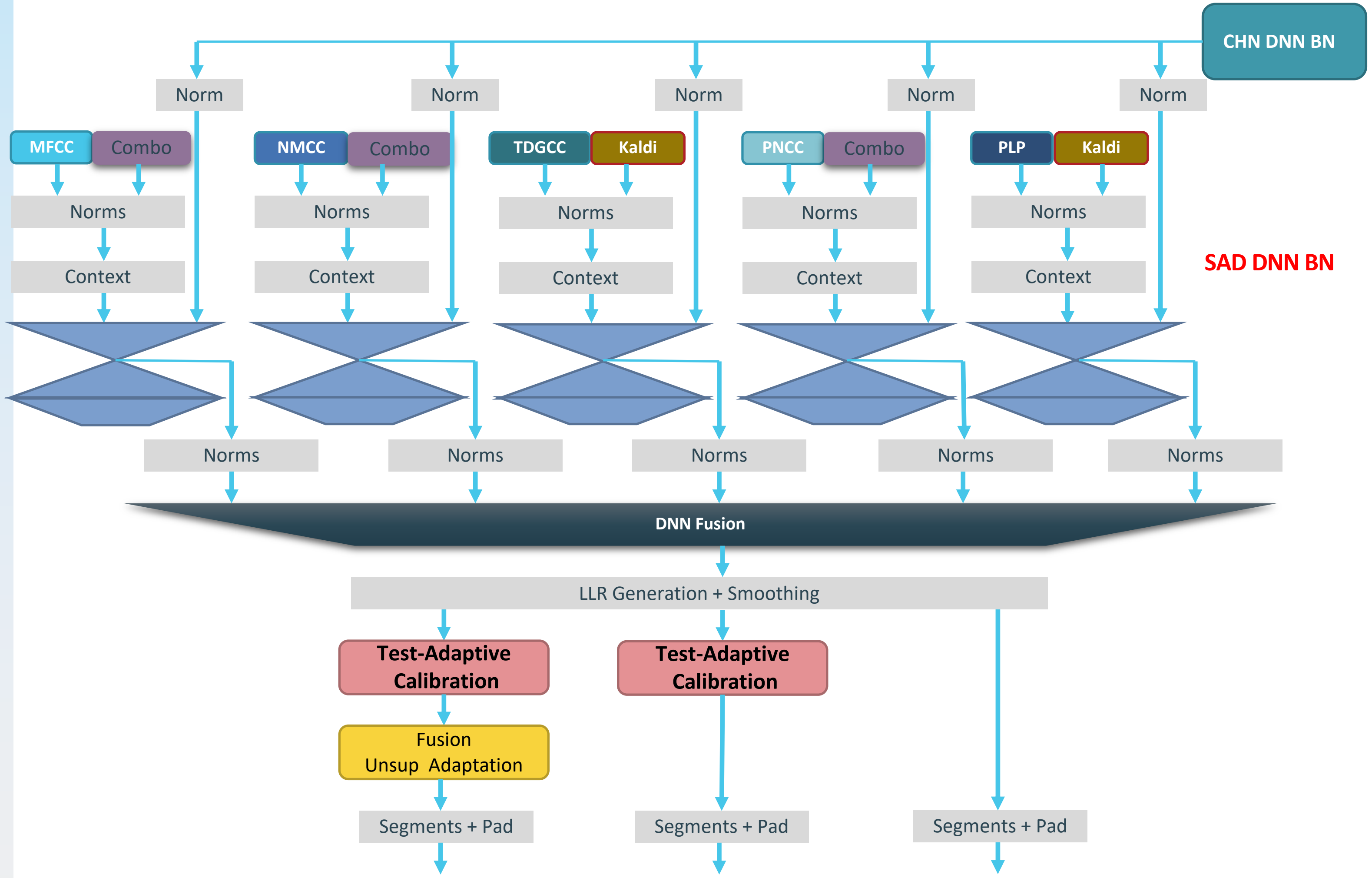


* Systems had different component parameters

SRI International

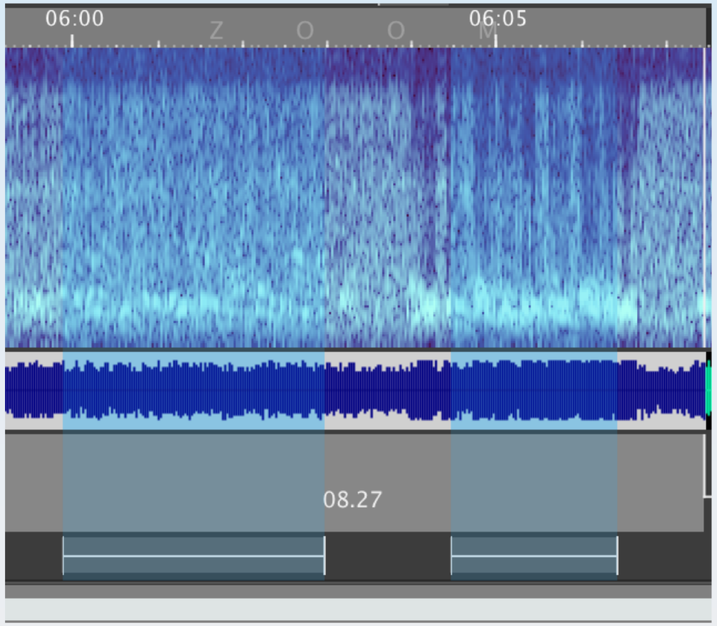
2015 DNN-based SAD

- Great performance on seen data after global calibration of fused system (over 10-20% improvement over GMM)
- More complex yet
- Feature combination at multiple levels
- Test-adaptive calibration crucial for generalizing to new data
- Unsupervised adaptation not always successful

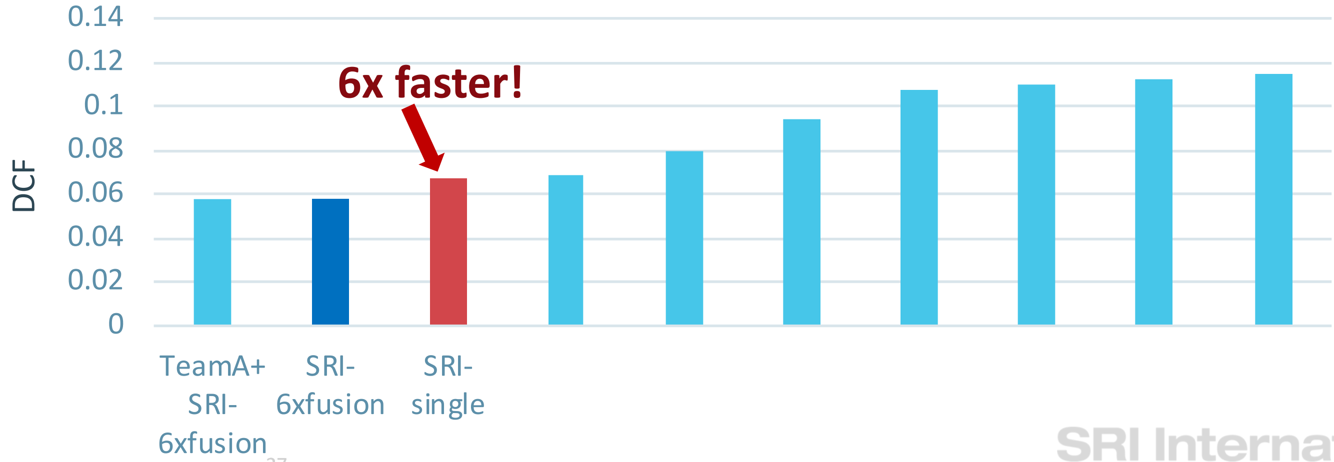


SRI International

SAD today



- Two important goals were not achieved with prior eval systems:
  - Generalization to unseen, highly variable, noisy conditions
  - High speed and low memory usage
- Developed single MFCC-based DNN system
  - Multi-condition training data
  - Feature normalization designed to minimize false alarms
- Applied simplified system in OpenSAT 2017
  - Comparable to best performing system on Video (VAST) track: real-world videos: spontaneous speech, background music, noises



Remaining SAD Challenges

For Data In-the-Wild

Adaptation and calibration to new conditions with

- Little data
- Unbalanced annotations
- Unsupervised data

Accuracy in the face of variable/unseen conditions

- Recent approaches explore use of LSTM models and acoustic embeddings

Often hard to draw line between speech and nonspeech

- Live human vs. TV/radio, intelligible speech vs unintelligible babble, singing

SRI International

Speaker ID

Goal: Identify a speaker known to the system, in potentially very different conditions from what it has seen

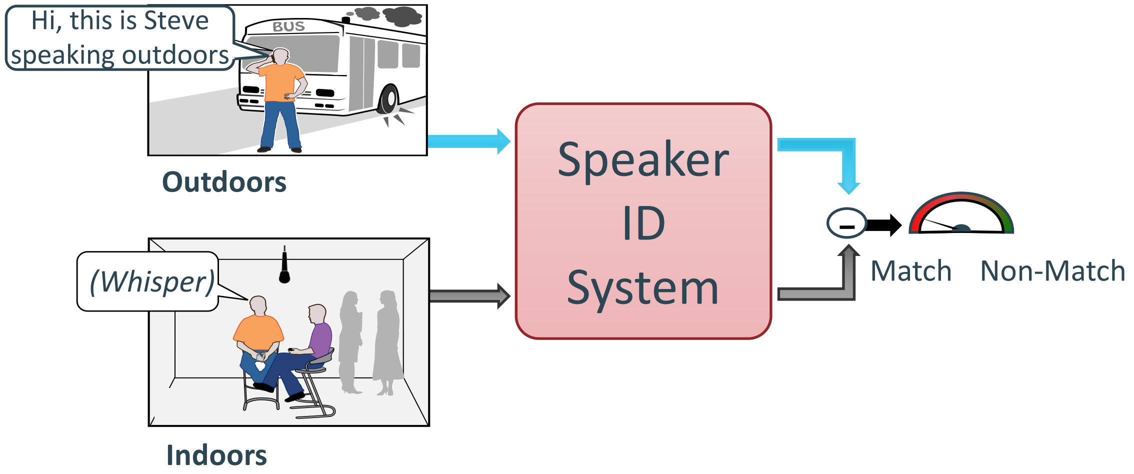
From UBM-based i-vectors and noise-robust features to deep learning and embeddings

Enrollment

- Get a sample of speech to model each speaker

Recognition

- Compare incoming speech to known speaker models



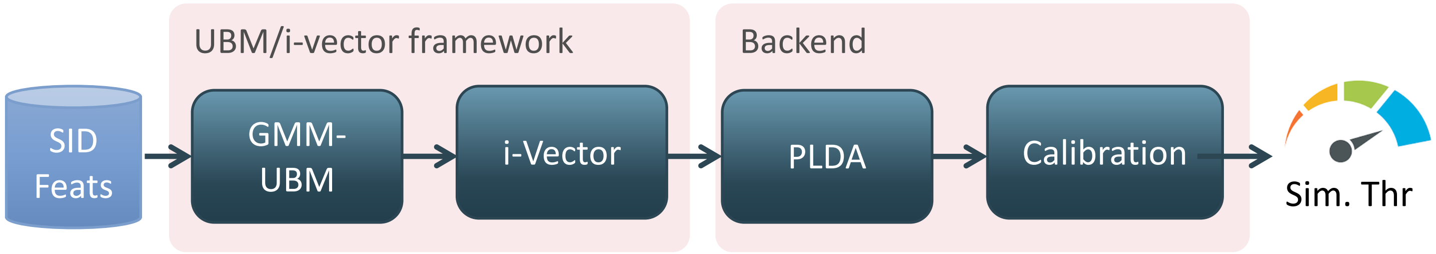
Challenging speech classification task since typically:

- there is very little training data for target classes (speakers)
- enrollment and test conditions for a speaker are mismatched

SRI International

Standard Approach 2012: UBM+ i-vectors + Bayesian Backend

Need effective and robust frontend representation and a modeling approach that can remove sources of variability



I-vectors: project variable-duration utterance onto a low-dimensional vector, typically of a few hundred components

Backend classification and calibration modules measure similarity to target speaker

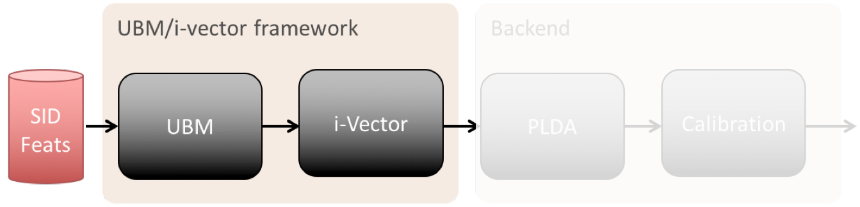
- PLDA: models speaker and intersession variability in the space of i-vectors, based on joint-factor analysis work

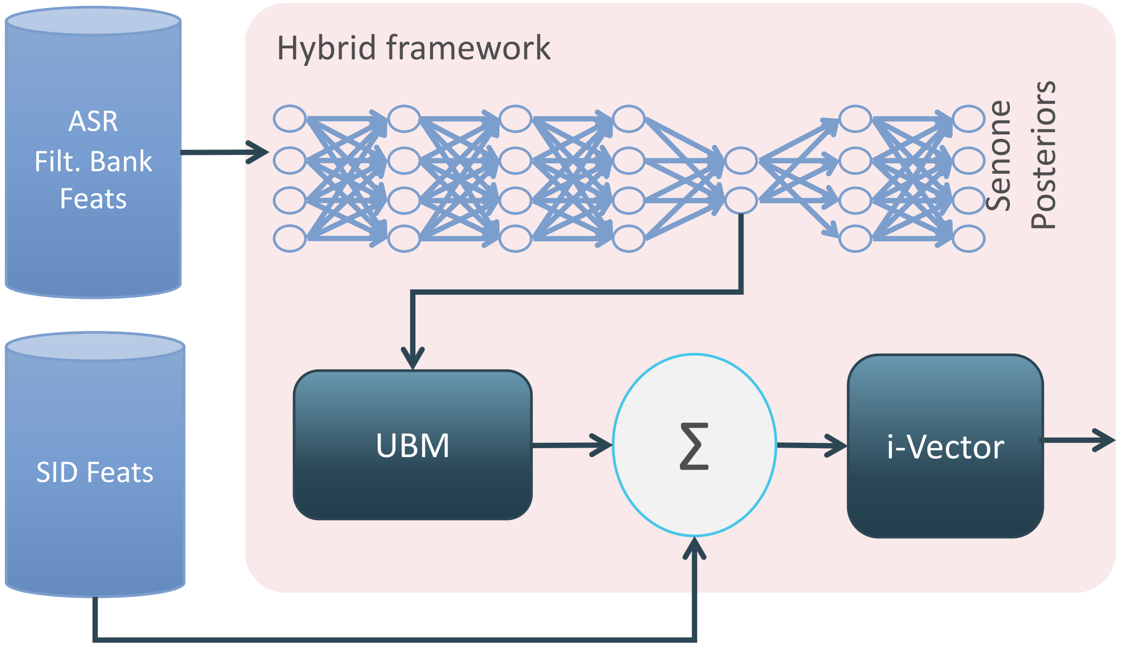
N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Frontend factor analysis for speaker verification," IEEE Trans. ASLP, vol. 19, May 2010

L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation", INTERSPEECH 2013

DNNs for SID: *Transfer learning for increased robustness*

- Replace GMM-based UBM with discriminative ASR-trained DNNs
  - UBM: Unsupervised sound clustering
  - DNN: Supervised, discriminative modelling of classes (senones)
- Use bottleneck rather than full senone posteriors
  - Lower dimensionality, faster computation
- Decouple frame alignment feature from SID feature
  - Reduce phonetic dependency in i-vectors

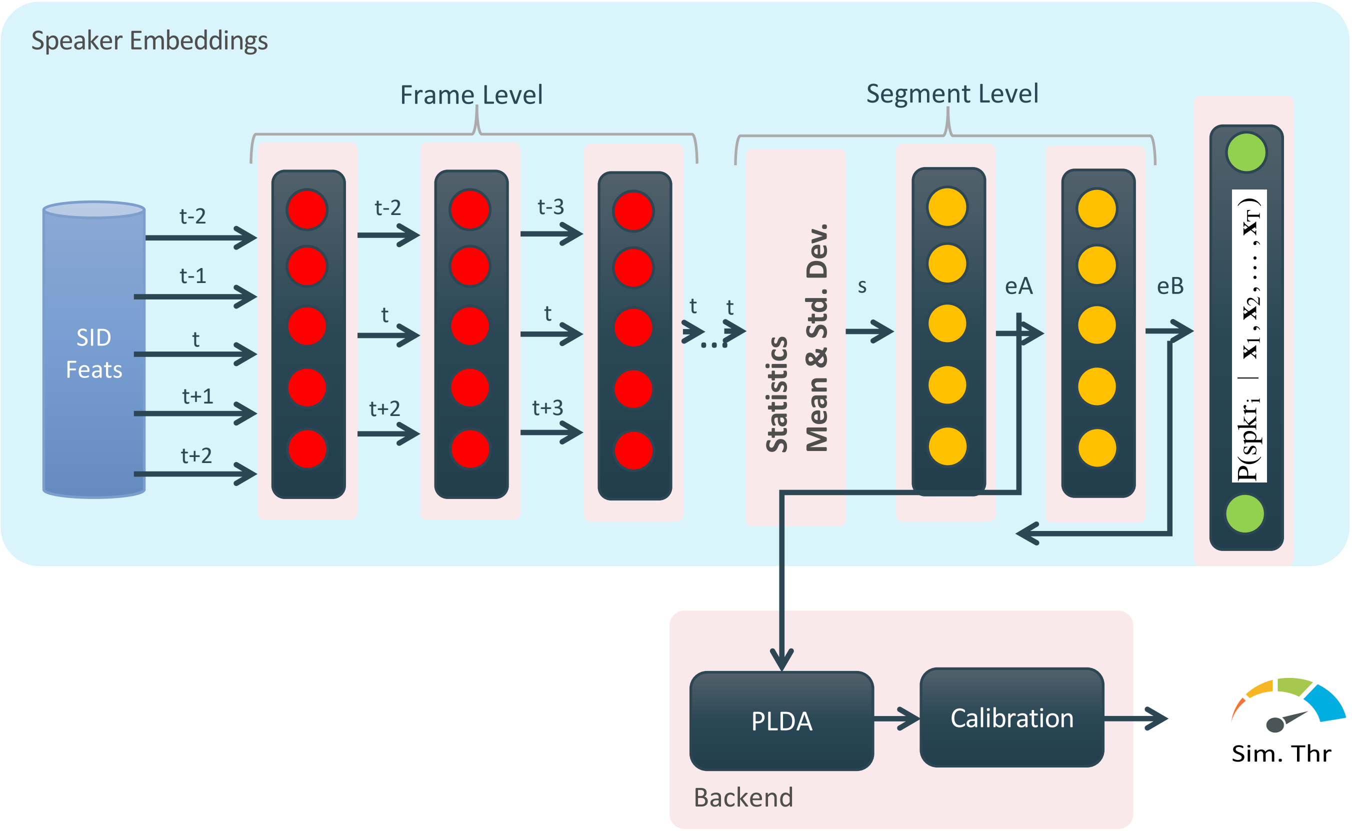




M. McLaren, D. Castan, L. Ferrer, A. Lawson, "On the Issue of Calibration in DNN-based Speaker Recognition Systems," in Proc. Interspeech 2016.

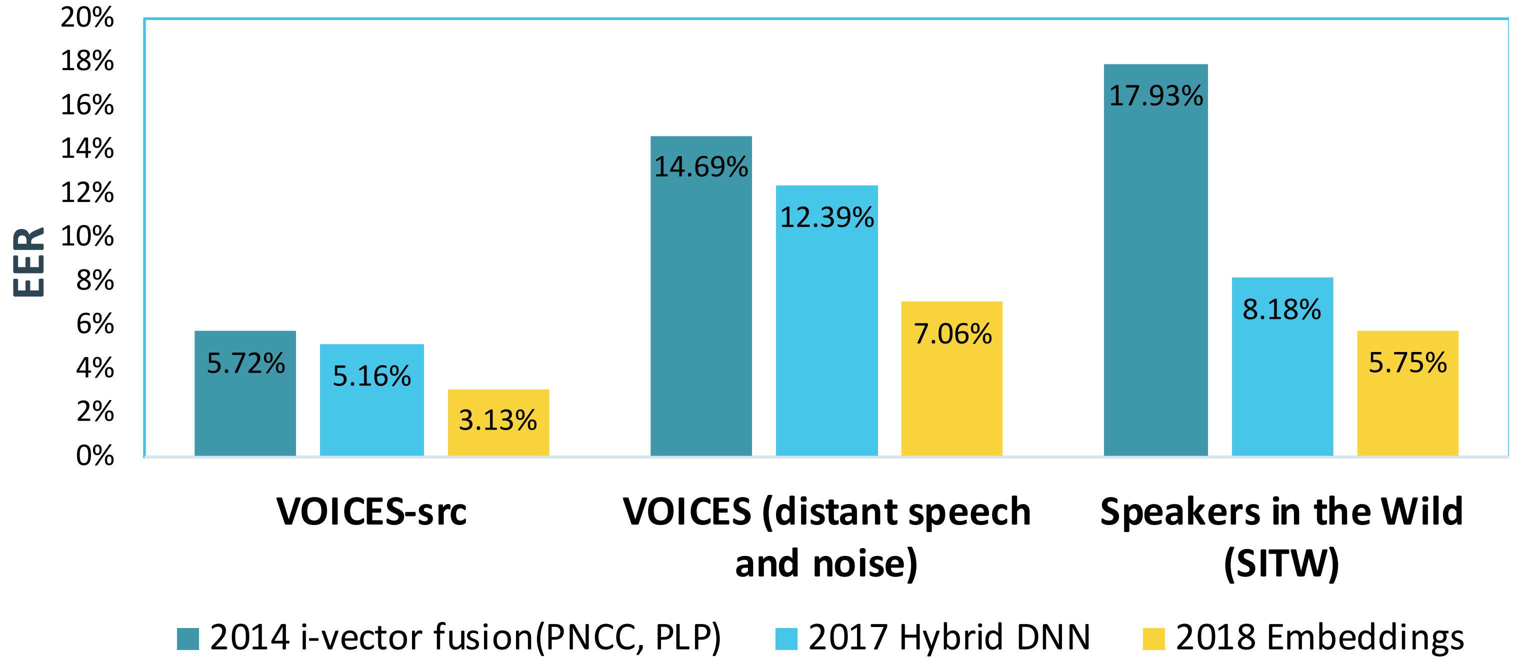
Low-dim representation with speaker embeddings (2017)

- Replace i-vectors with embeddings extracted from a feed-forward DNN
- Long-term speaker characteristics are captured by a temporal pooling layer
  - Mean and standard deviation on the segment (2 - 10s)



D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," ICASSP, 2018

Benchmarking progress in realistic conditions

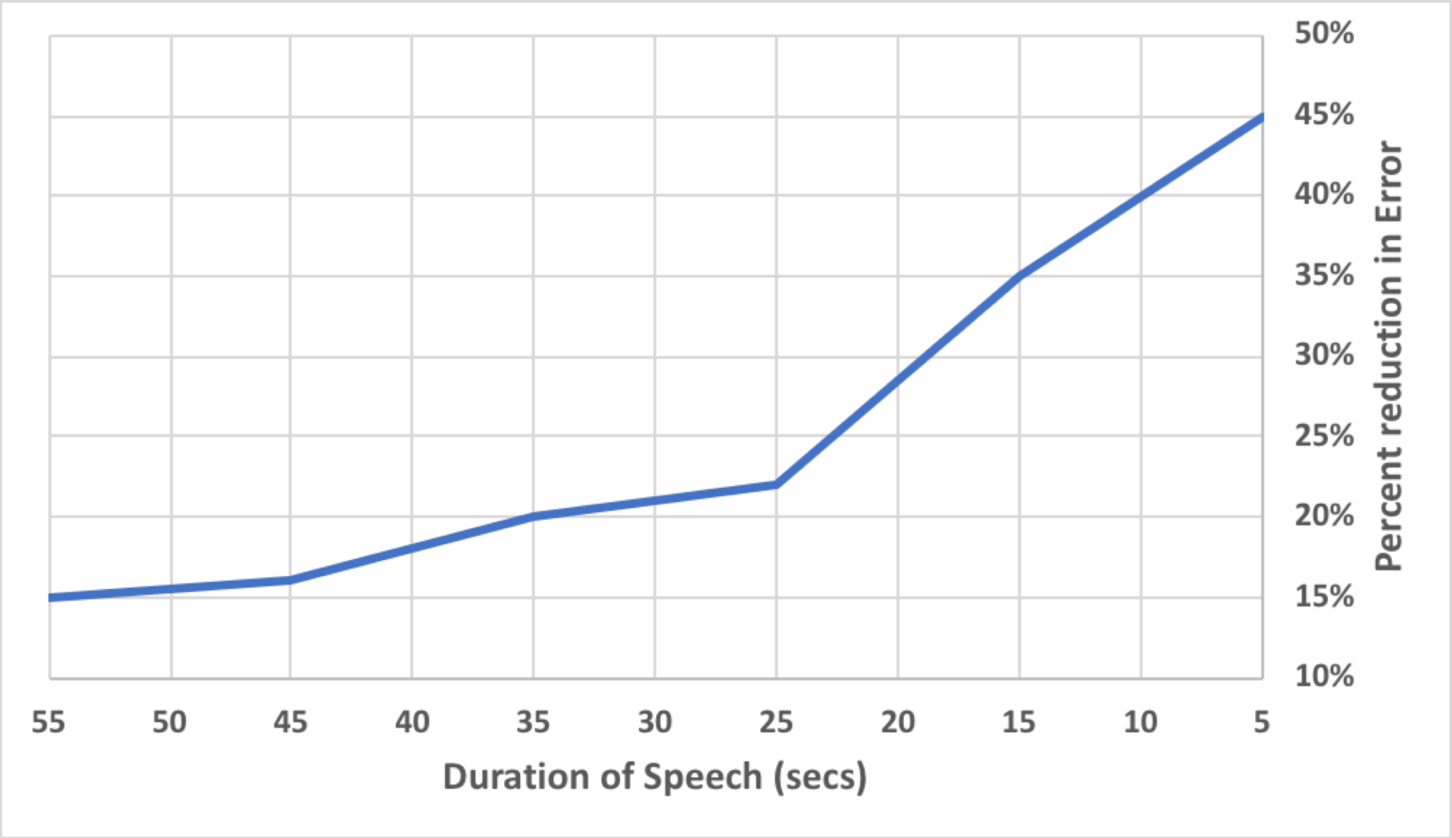


M. K. Nandwana, J. van Hout, M. McLaren, A. Stauffer, C. Richey, A. Lawson, M. Graciarena, "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings," in ISCA INTERSPEECH 2018

Effectiveness of embeddings in short durations

Current SID approach works as well on 2 sec audio segments as pre-embeddings did on 8 sec

At 5 sec duration, embeddings reduce error by 45%



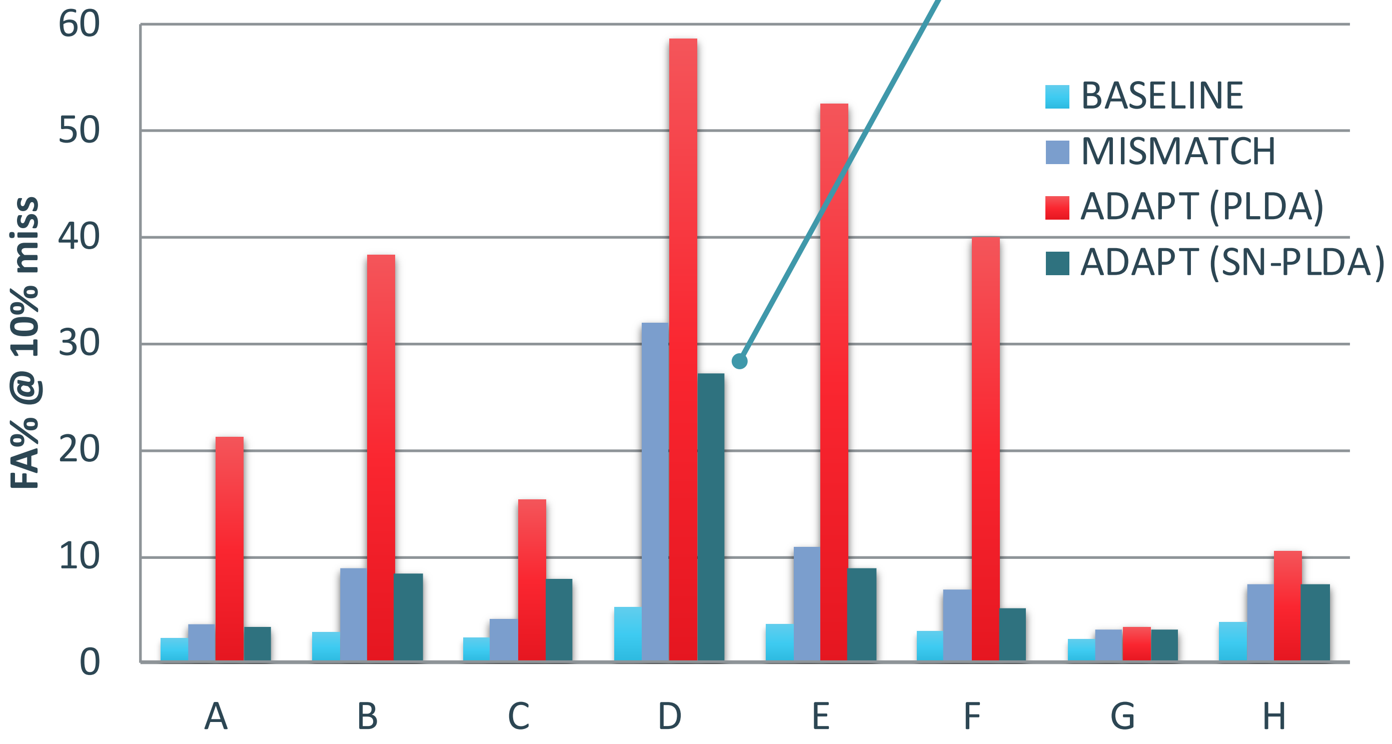
SRI International

Need for Field Adaptation

- Big degradation in mismatched conditions (from RATS data)
- Simply adding data to retrain PLDA does NOT work due to underlying assumptions on speakers' distributions among different conditions across original and adaptation data
- Source-normalization makes adaptation data compatible to original model

Source Normalization (SN)
enables gain from adaptation data

Further research is needed to better exploit the limited adaptation data

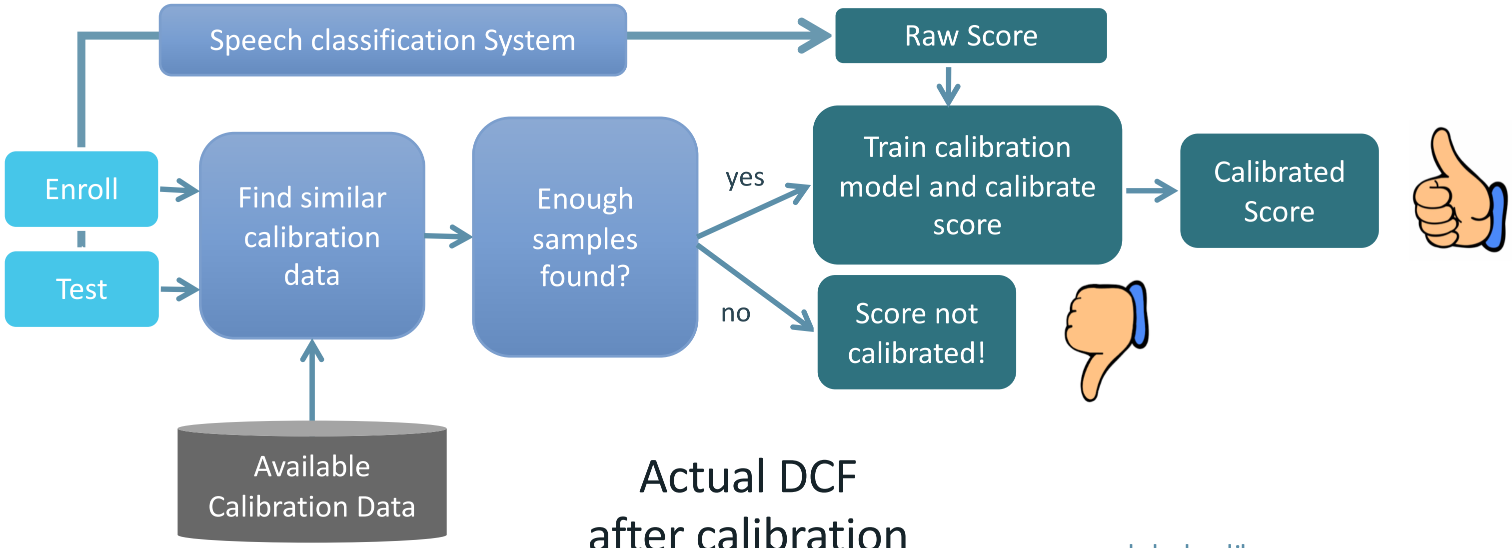


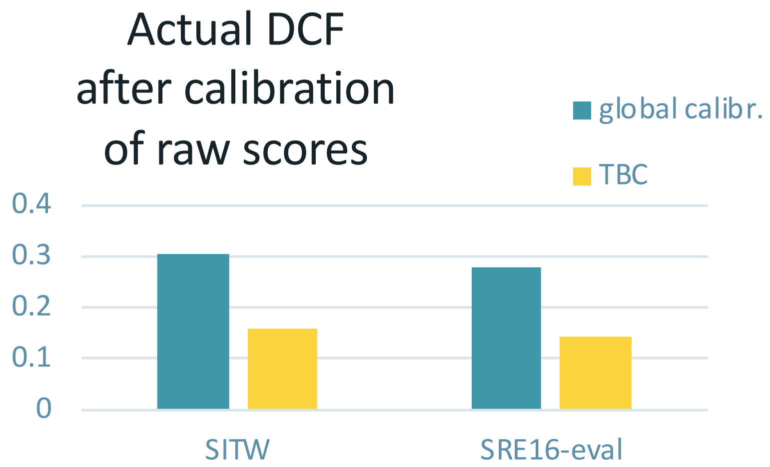
M. McLaren, and D. Van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," Audio, Speech, and Language Processing, IEEE Transactions on 20 (3), 755-766, 2012

SRI International

Trial-Based Calibration (TBC): Towards fail-safe calibration

- Key to good calibration: use data that properly represents the trial conditions
- In the wild, conditions are different for every trial
- Relevant trials are found with a metric of similarity between the acoustic conditions of two signals
- When not enough trials are selected to train a calibration model, the trial is not calibrated
  - Better not to calibrate at all then calibrate badly





M. McLaren, A. Lawson, A., L. Ferrer, N. Scheffer, and Y. Lei, "Trial-based calibration for speaker recognition in unseen conditions", In Odyssey 2014

L. Ferrer, M. K. Nandwana, M. McLaren, D. Castan, and A. Lawson. "Toward Fail-Safe Speaker Recognition: Trial-Based Calibration With a Reject Option". IEEE/ACM Transactions on Audio, Speech, and Language Processing 27, no. 1 (2019): 140-153

Remaining SID Challenges

Multi-speaker SID

- Data in the wild is hardly ever single speaker

Intrinsic speaker variability

- Impact by emotion, stress, vocal effort, etc.

Optimal speaker embedding computation

M, McLaren, D. Castan, M. Kumar Nandwana, L. Ferrer and E. Yilmaz. "How to train your speaker embedding extractor" Speaker Odyssey 2018

Unsupervised adaptation to new conditions

- Trial based calibration is one approach

Confidence and calibration

- When do we trust the system output?
- How to avoid miscalibration

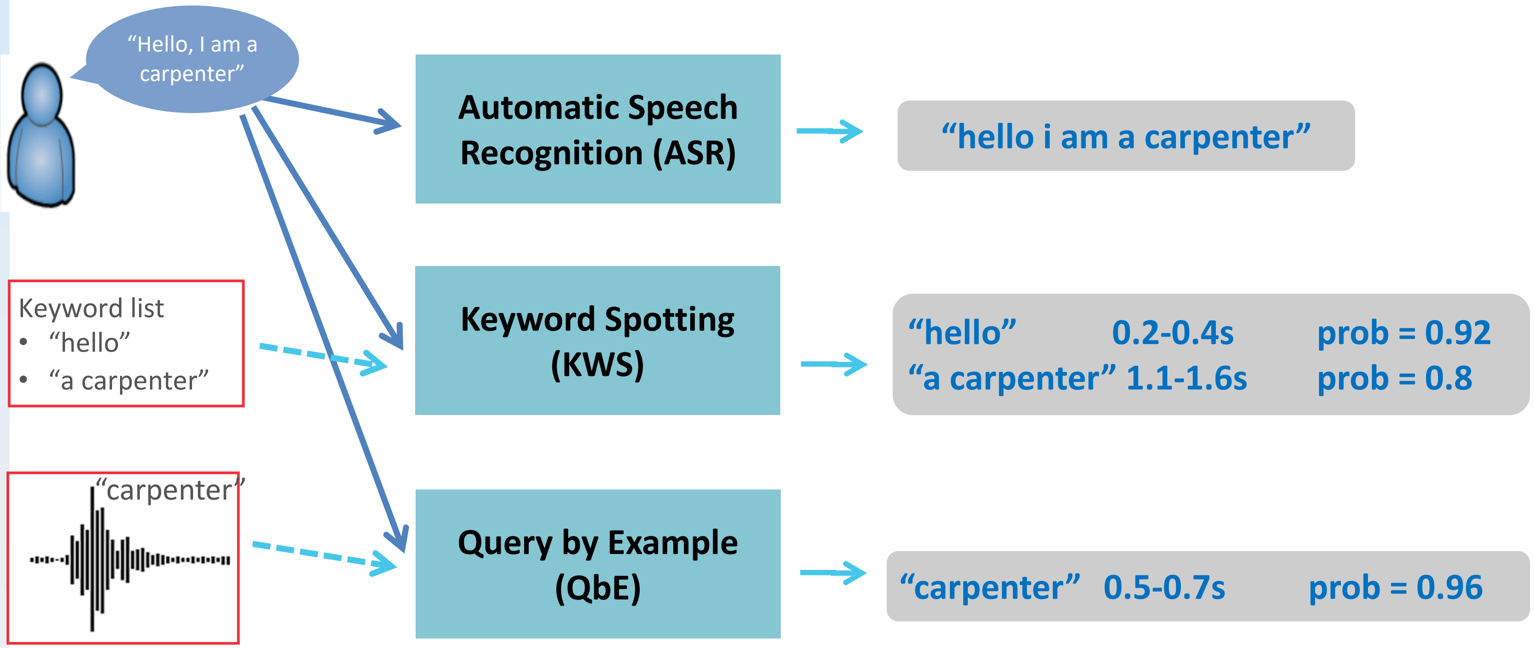
SRI International

Query-by-Example (QbE)

Robust keyword detection under very challenging acoustic conditions

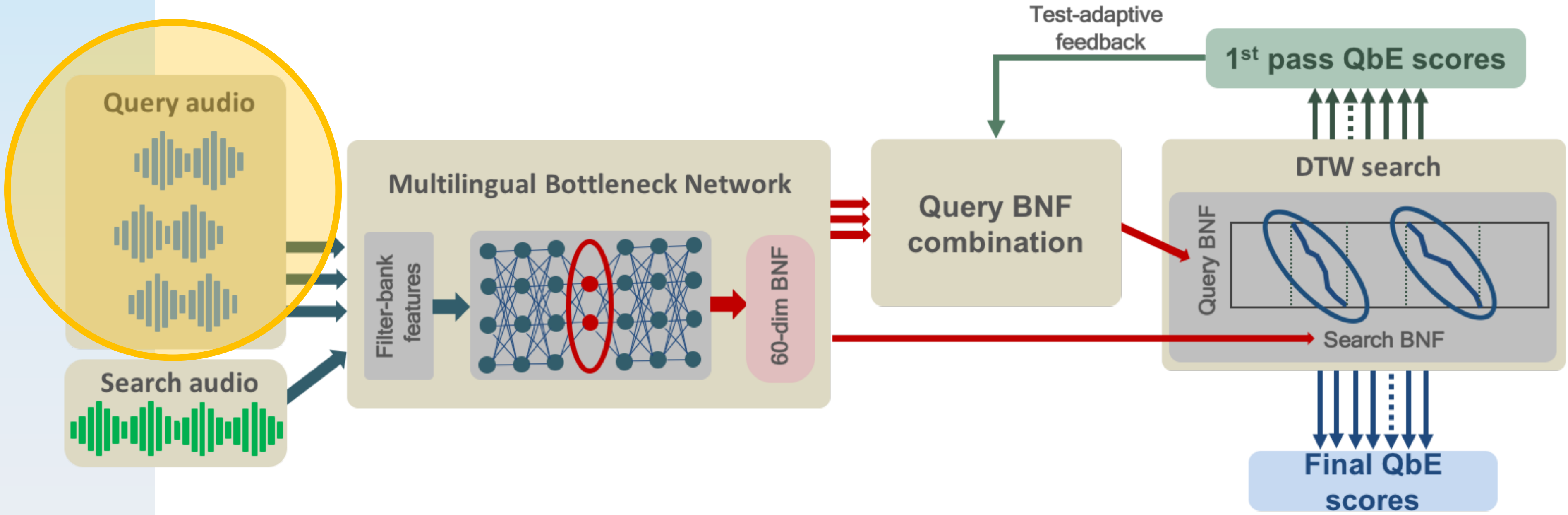
Keyword Spotting and Query-by-Example: Finding Important Spoken Words

- When ASR is not accurate enough, focus only on keywords
- Keyword Spotting (**KWS**) finds word probabilities **using the most likely word sequences** from ASR
  - Degrades for OOVs or for low ASR performance
- Query-by-Example (**QbE**) lets the user **select an audio sample** and search for other occurrences of that keyword
  - Language independent
  - One-shot learning



SRI International

QbE DNN-based Implementation

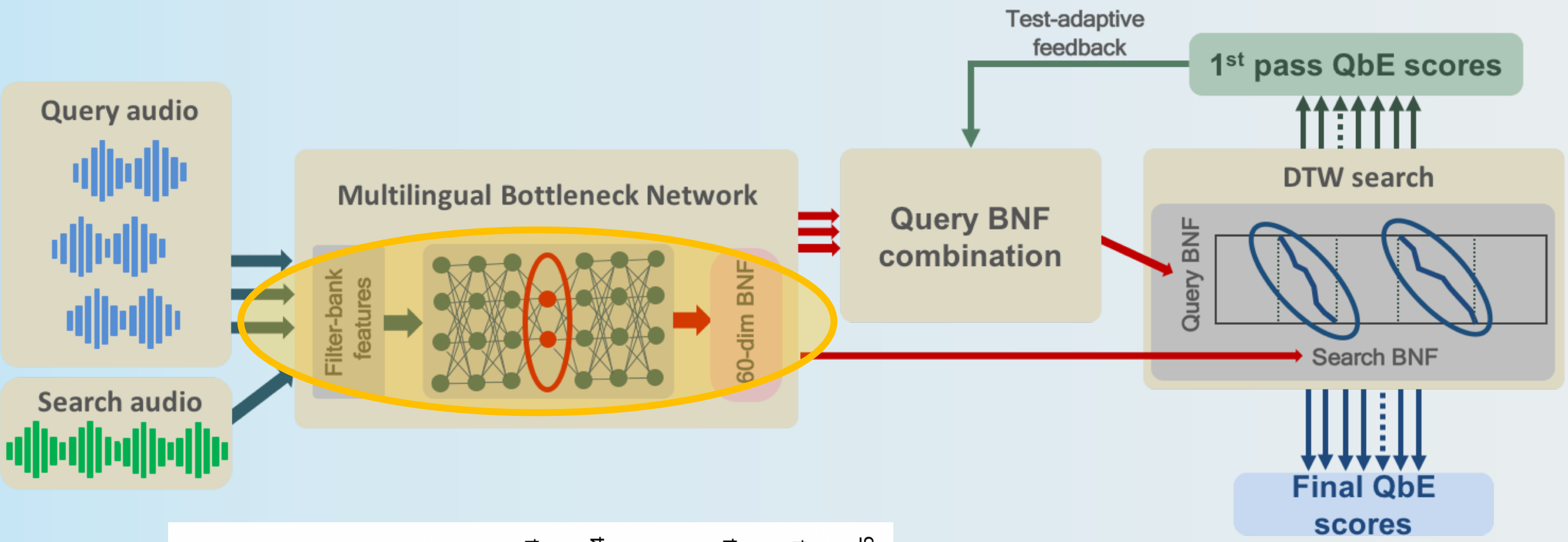


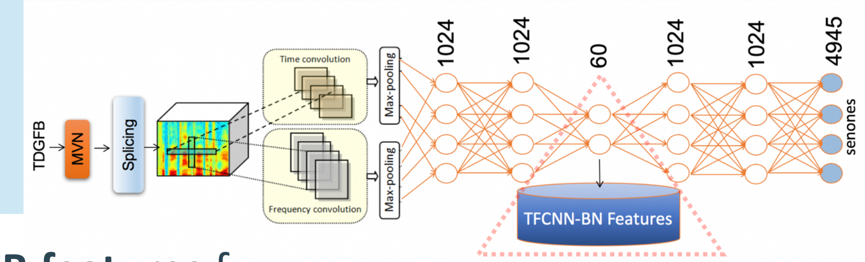
E. Yılmaz, J. van Hout and H. Franco. "Noise-Robust Exemplar Matching for Rescoring Query-by-Example Search." IEEE ASRU 2017

Noise robust SAD is needed to remove silence and noise from the query

QbE DNN-based Implementation

Word embedding representation





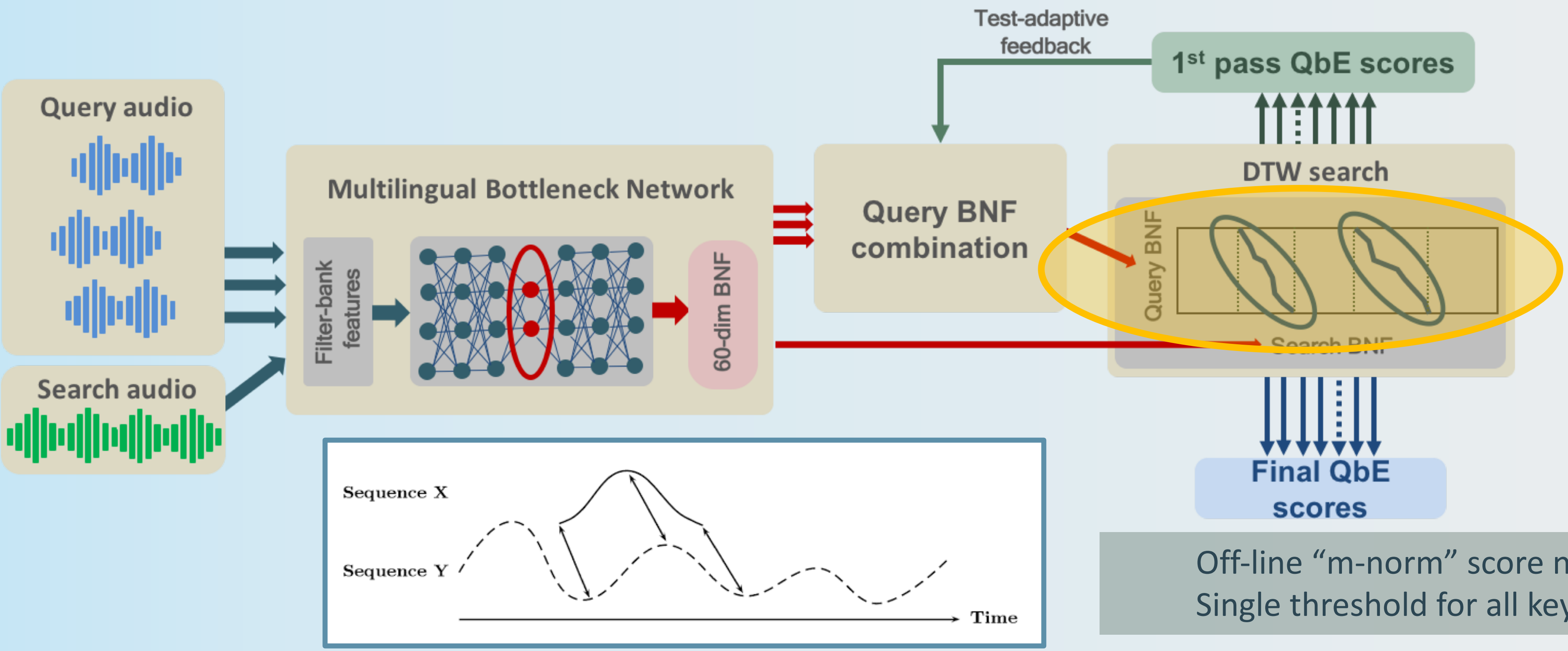
Trained **on multi-lingual data set**

Time-domain GFB features for noise robustness

SRI International

QbE DNN-based Implementation

Search



Off-line "m-norm" score normalization
Single threshold for all keywords

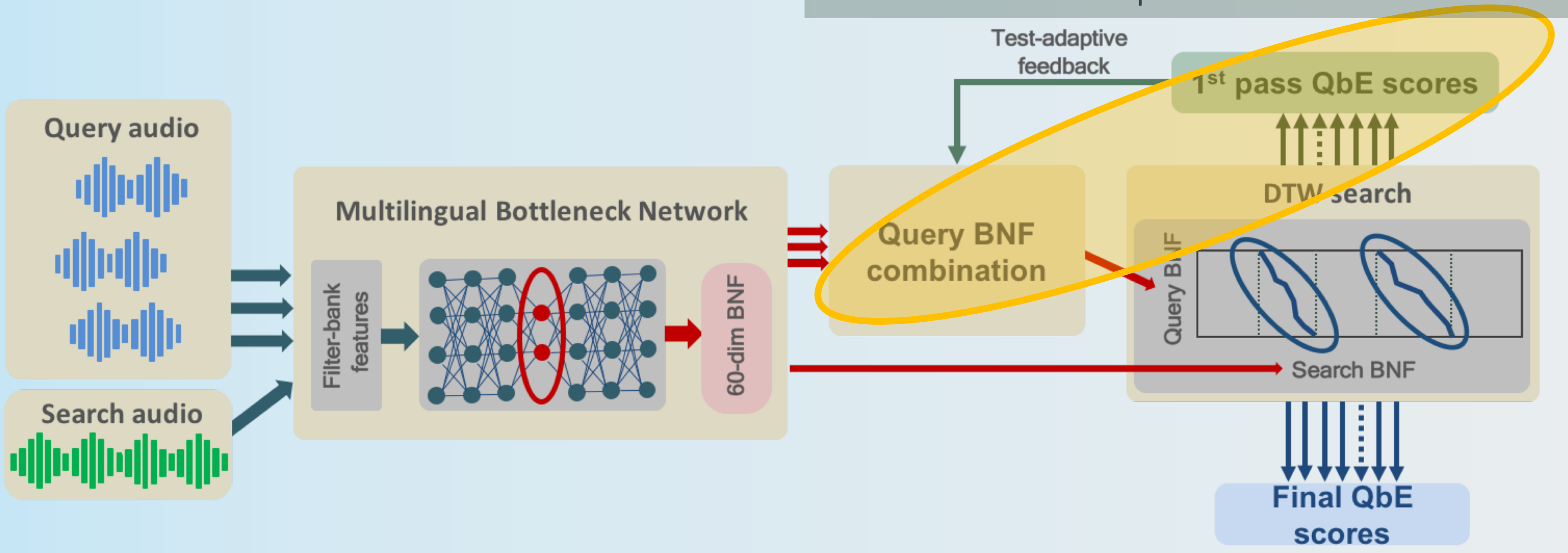
Sub-sequence Dynamic Time Warping
finds optimal query/utterance alignments

SRI International

QbE DNN-based Implementation

Unsupervised optimization of weights for the case of multiple examples

- Iterative examples alignment
- 1st pass BNF averaging produce meta-example
- Gradient descent to pick detection-specific weights
- Combine with 1st pass results



SRI International

Application Domain: Air Traffic Control

Realistic and challenging conditions

- Highly degraded acoustics: English conversations recorded from the ground from planes and control towers
- Acoustic mismatch in enrollment and test: various airports, controllers, planes,..
- Enrollment from keywords pronounced mid-sentence, more realistic

Sample keyword 1

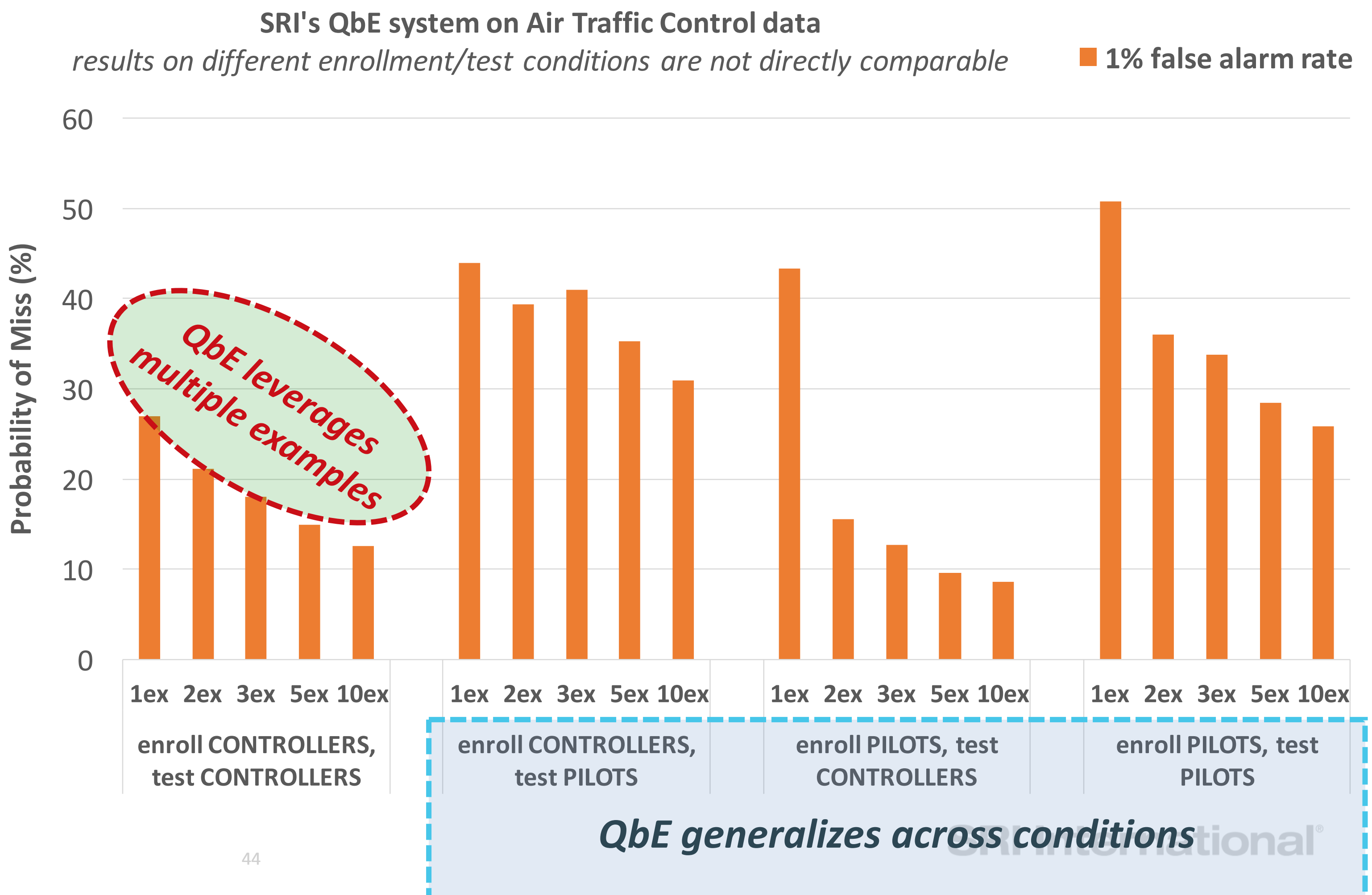
plane tower

"runway two two right"

Sample keyword 2

plane tower

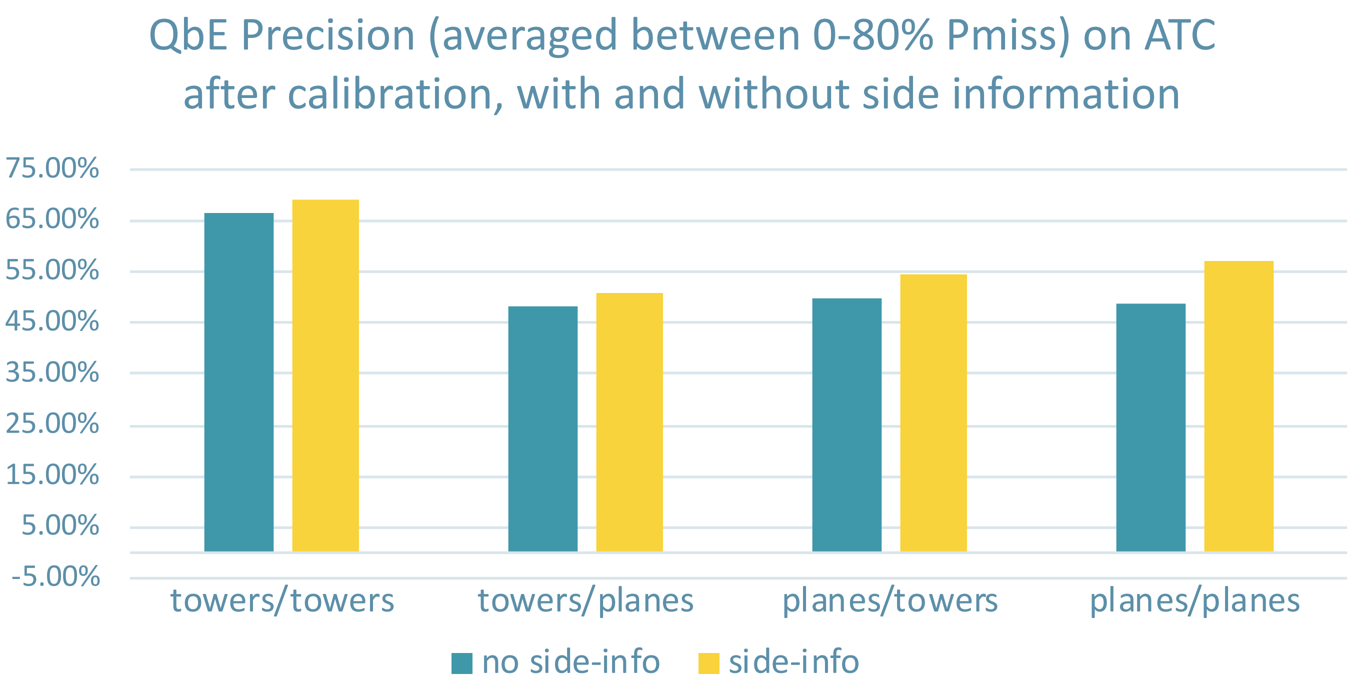
"precision thirty-five"

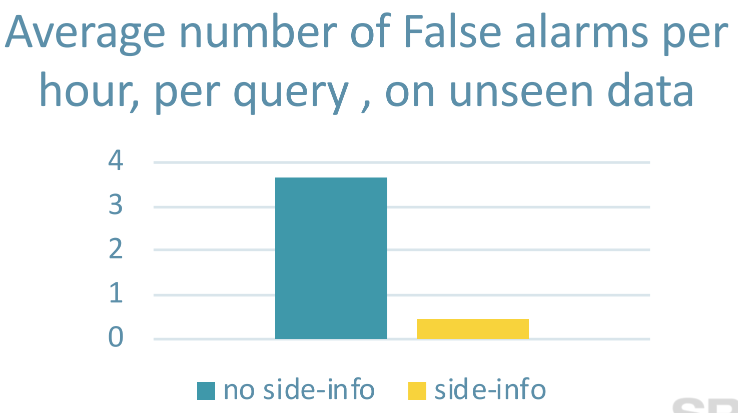


Score calibration using side info to reduce false alarms

Use detections to self-calibrate and avoid over-matching when using multiple examples

- Optimize on mismatched data (i.e. enroll planes/test towers, or enroll towers/test planes) for generalization





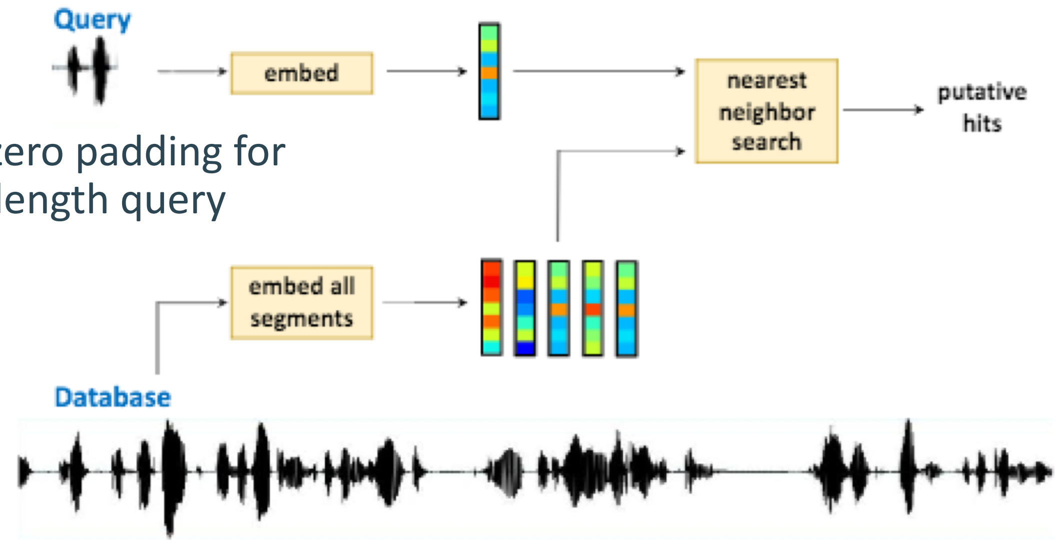
SRI International

QbE using Acoustic Word Embeddings

Replacing BN features vectors with Word Embeddings results in replacing DTW with simple cosine distance

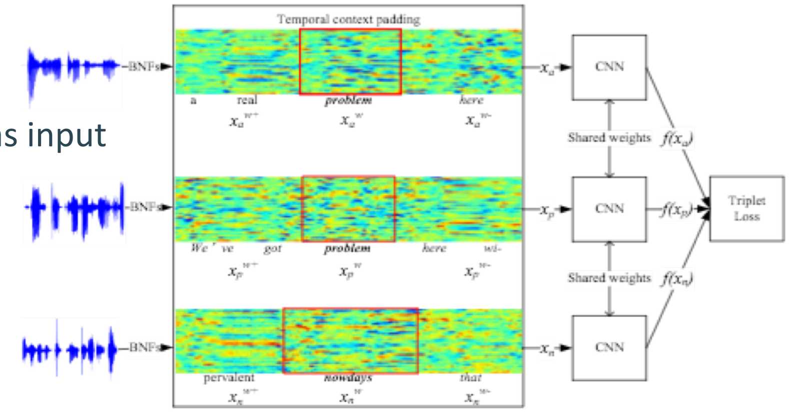
S. Settle, K. Levin, H. Kamper, and K. Livescu, "Query-by-example search with discriminative neural acoustic word embeddings", InterSpeech, 2017.

Y. Yuan, C. Leung, L. Xie, H. Chen, B. Ma and H. Li, "Learning Acoustic Word Embeddings with Temporal Context for Query-by-Example Speech Search", InterSpeech, 2018



uses zero padding for fixed length query

uses BN features as input



In Yuan et al, embeddings are learned via CNNs, with a triplet loss: uses two examples of same word (target), and one negative example.

Remaining QbE Challenges

Improve robustness of DNN-BNF / acoustic embedding features to handle most challenging acoustic conditions:

- HF/VHF radio distortions
- Bursts of noisy events
- Distant speech

Balance discrimination with generalization:

- Avoid confusions between similar phrases
  - ‘Big black car’ vs ‘big black bear’
- Enable inexact matching for applications in morphologically rich languages

Conclusions and Future Directions

Conclusions

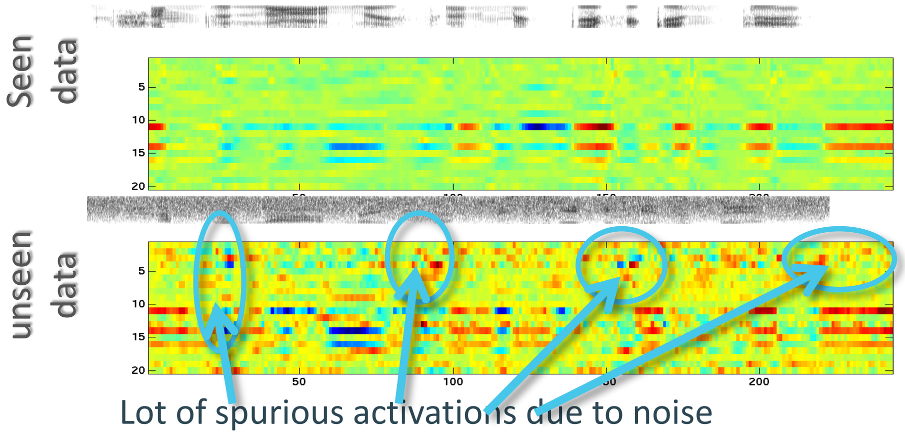
- Deep learning approaches have led to big improvements for all speech analytics tasks reviewed
- Handling unseen conditions is still a challenge
  - Embeddings and bottleneck features improve performance substantially
    - They are learned from data and represent what they have seen well
    - Noise robust features have been replaced, but there may be benefit in using them for increased robustness of embeddings' computation
- Score Calibration is crucial for performance and interpretability
  - Adaptive calibration helps for conditions in-the-wild
  - Calibration sensitive to proper data selection for the task

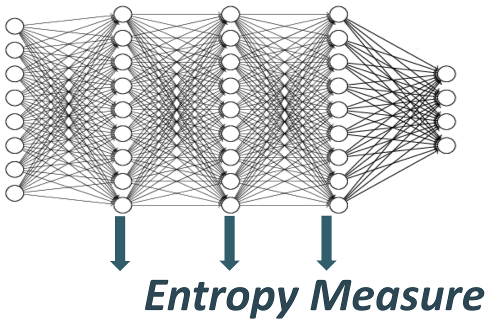
- Confidence & Interpretability: When and why can a system output be trusted?

SRI International

Potential Direction: Peeking into the DNN Activations

- For unseen data, DNN activations can be extremely noisy
- Extraction of a run-time activation entropy can provide some measure of DNN decision confidence
  1. Can we leverage this to predict when the DNN is witnessing unseen data?
  2. Can we use this information to select data for adaptation and calibration?





V. Mitra and H. Franco, "Interpreting DNN output layer Activations: A strategy to cope with Unseen Data in Speech Recognition," in Proc. of ICASSP 2018.
V. Mitra, H. Franco, C. Bartels, J. van Hout, M. Graciarena and D. Vergyri, "Speech Recognition In Unseen And Noisy Channel Conditions," in Proc. of ICASSP 2017

SRI International

Thank you!

2018 IEEE Spoken Language Technology Workshop

SRI International:

Independent nonprofit research center in Silicon Valley, founded in 1946 by Stanford University

Core business: Science and technology solutions for government and businesses worldwide

- Basic research
- Systems and solutions
- Venture incubation
- Technology licensing

4000+ patents

70+ spin offs

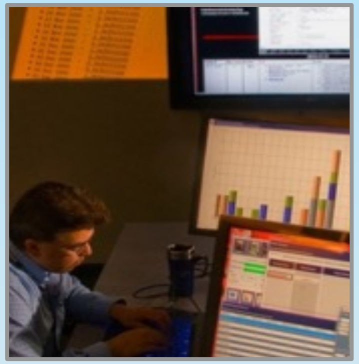
1,700 employees, **30** labs, **21** locations

Now hiring: https://www.sri.com/careers

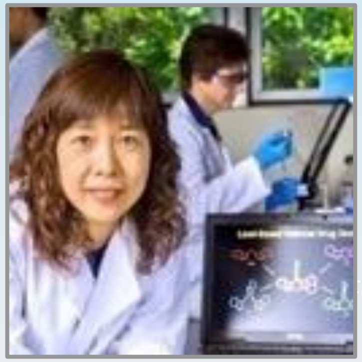
First computer mouse

First ARPANET and internetworking nodes

Electronic Banking



Cyber Security



1st drug for malaria; drugs for lymphoma

Dept. of Education 2010 tech plan

First telerobotic surgical system

Ultrasound for Medical diagnostics

NUANCE

(1994)

(acquired by Apple in 2008)

Kasisto

(2014)