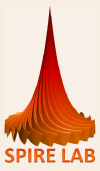
A Comparative Study on the Effect of Different Codecs on Speech Recognition Accuracy Using Various Acoustic Modeling Techniques

Srinivasa Raghavan[1], Nisha Meenakshi G[1], Sanjeev Kumar Mittal[1], Chiranjeevi Yarra[1], Anupam Mandal[2], K.R. Prasanna Kumar[2], Prasanta Kumar Ghosh[1]

[1]SPIRE LAB, Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India,
[2]Center for AI and Robotics, Bangalore, Karnataka, India

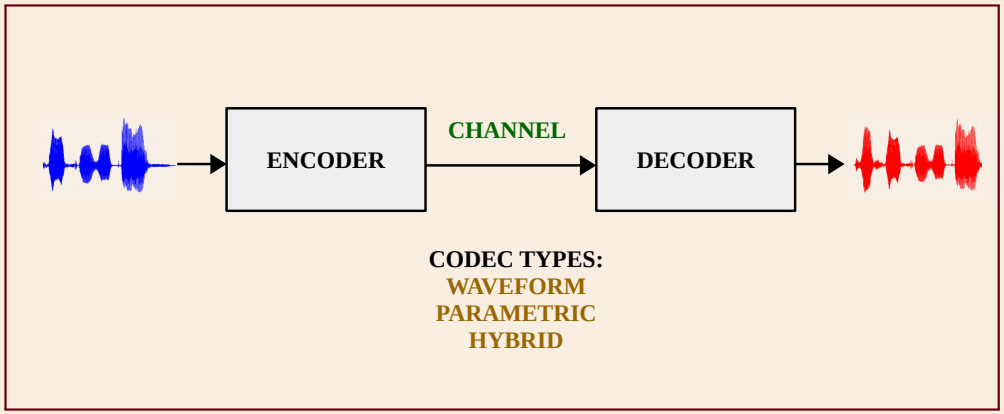


Section 1

Focus

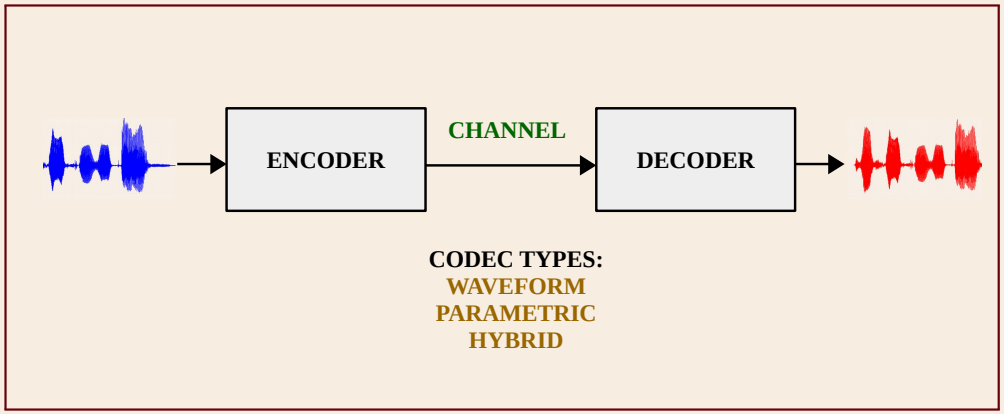
A Comparative Study on the Effect of Different **Codecs** on **Speech Recognition** Accuracy Using Various Acoustic Modeling Techniques.

Speech Coding & Automatic Speech Recognition (ASR)

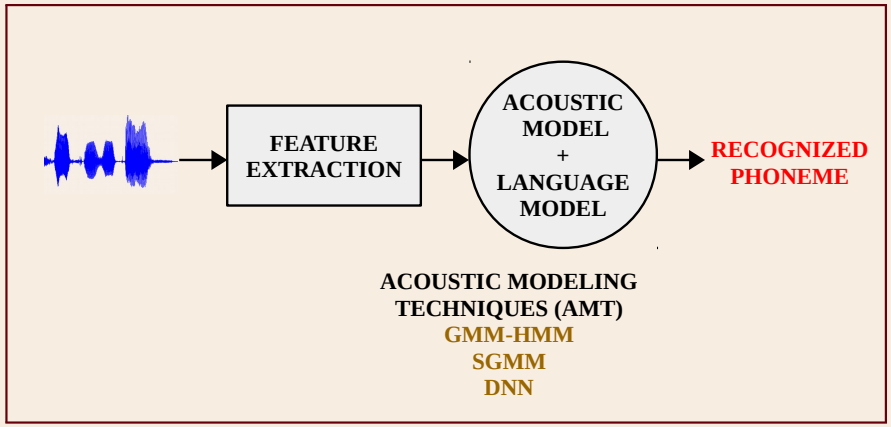


Speech Coding

Speech Coding & Automatic Speech Recognition (ASR)

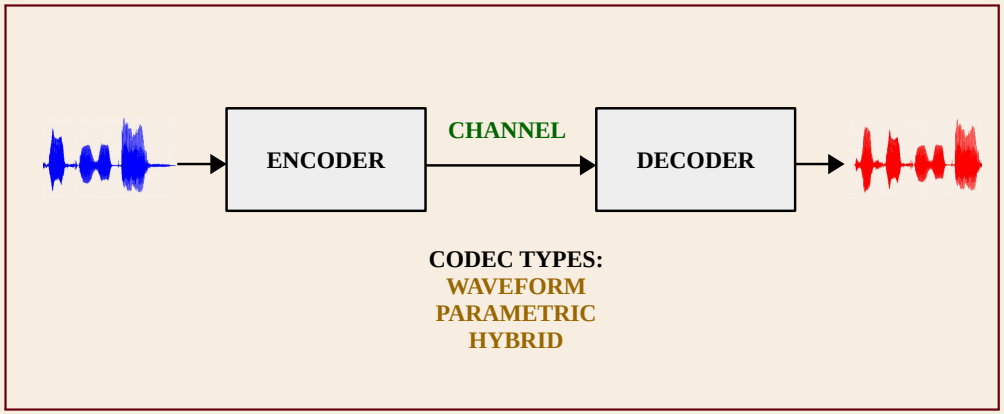


Speech Coding

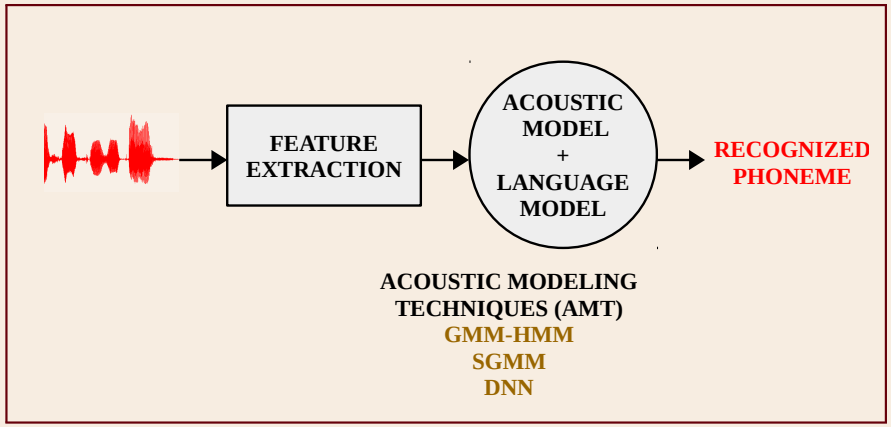


Speech Recognition

Speech Coding & Automatic Speech Recognition (ASR)

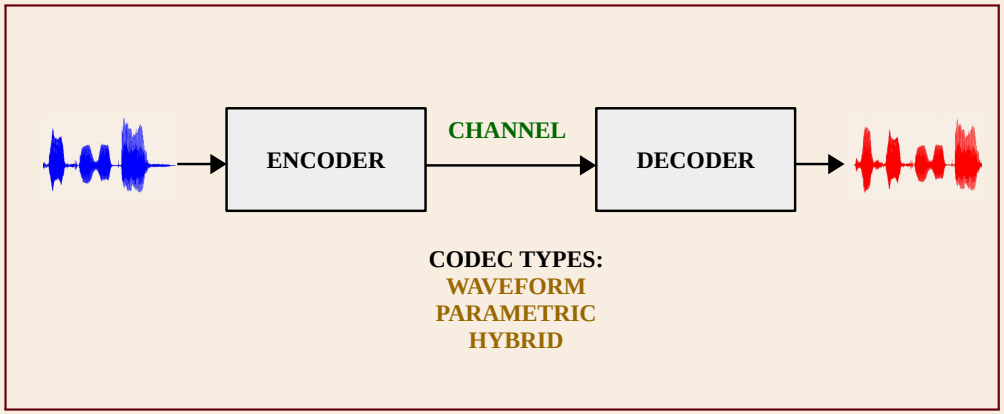


Speech Coding

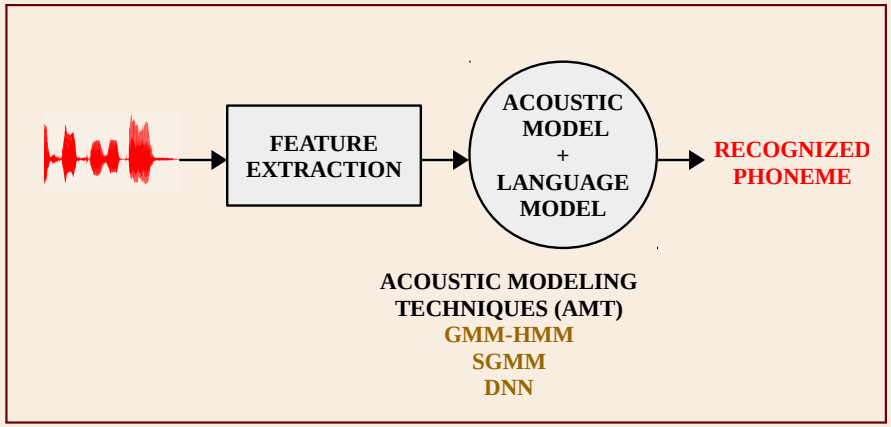


ASR with Codec Distorted Input

Speech Coding & Automatic Speech Recognition (ASR)



Speech Coding

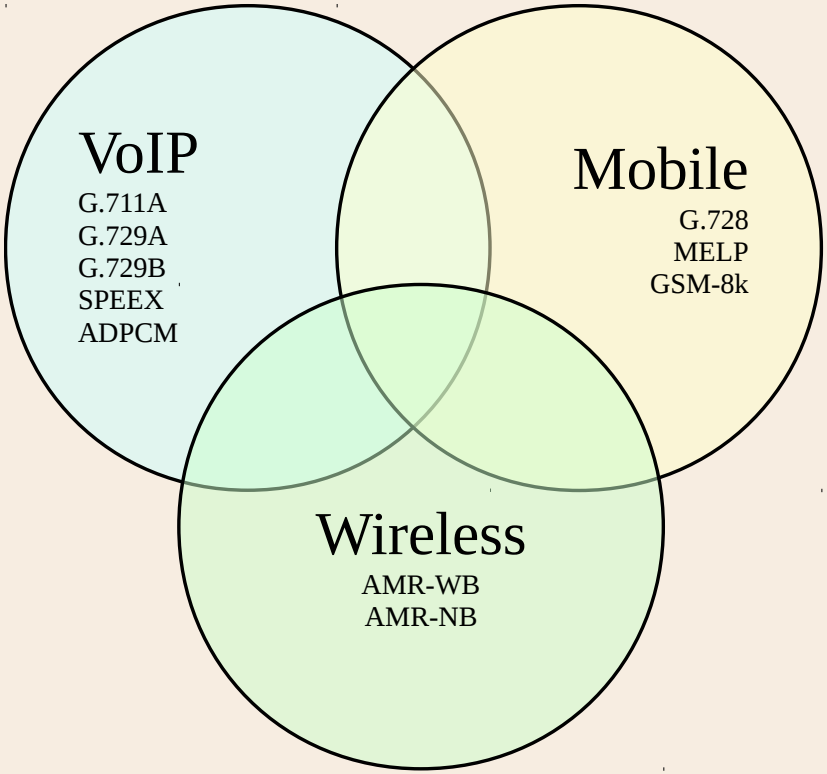


ASR with Codec Distorted Input

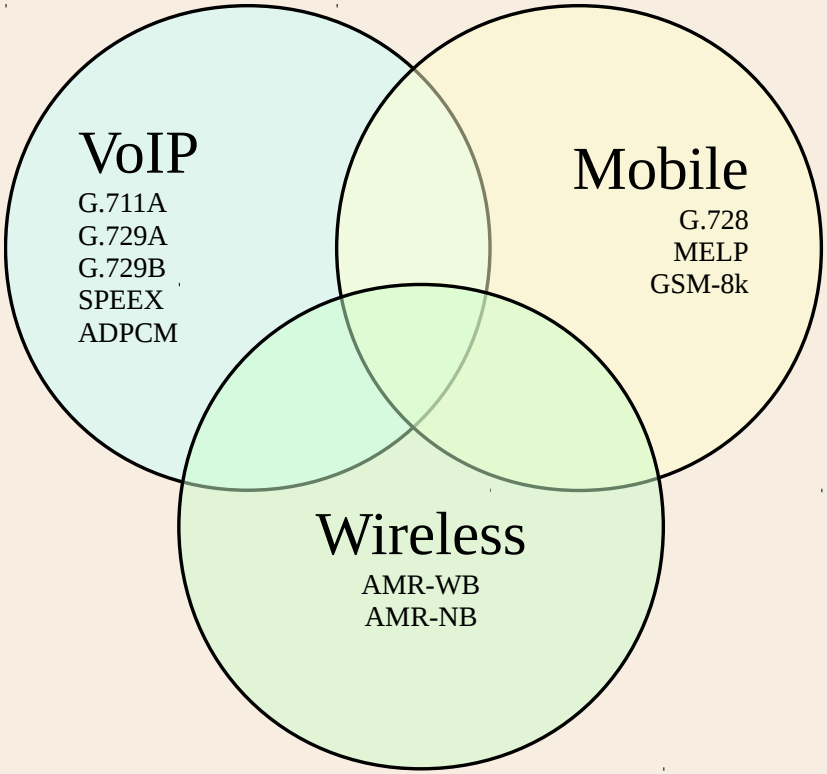
Note

1. The **Channel Effect** is not considered.
2. Effect of **Language Model** is not considered.

Common Speech Coders

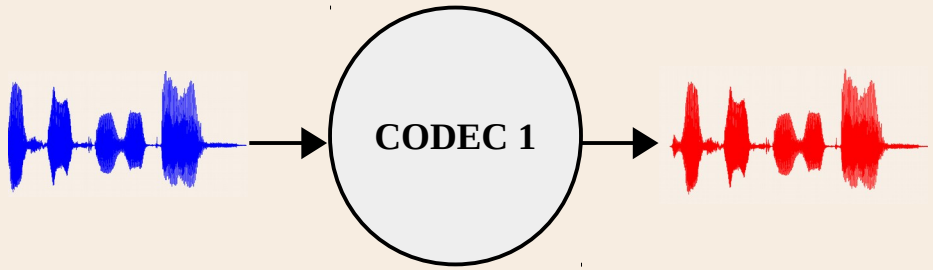


Common Speech Coders



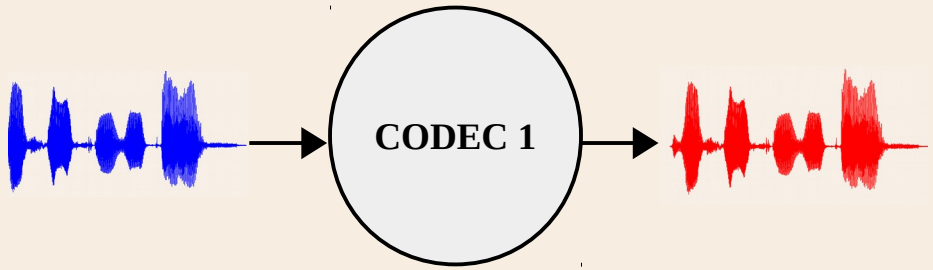
Codec	Type	Band-width	Bit-rate (kbps)
G.711A	Waveform	Narrow	64
MELP	Parametric	Narrow	2.4
AMR-NB	Hybrid	Narrow	4.40
AMR-WB	Hybrid	Wide	23.85
G.728	Hybrid	Narrow	16
G.729A	Hybrid	Narrow	8
G.729B	Hybrid	Narrow	8
PCM	Waveform	Narrow	128
ADPCM	Waveform	Wide	32
GSM-8k	Hybrid	Narrow	13
SPEEX	Hybrid	Wide	27.8

Common Speech Coding Strategies

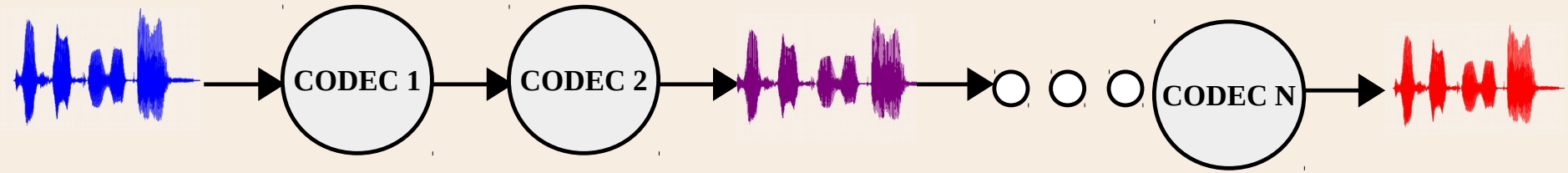


Single Encoding-Decoding

Common Speech Coding Strategies

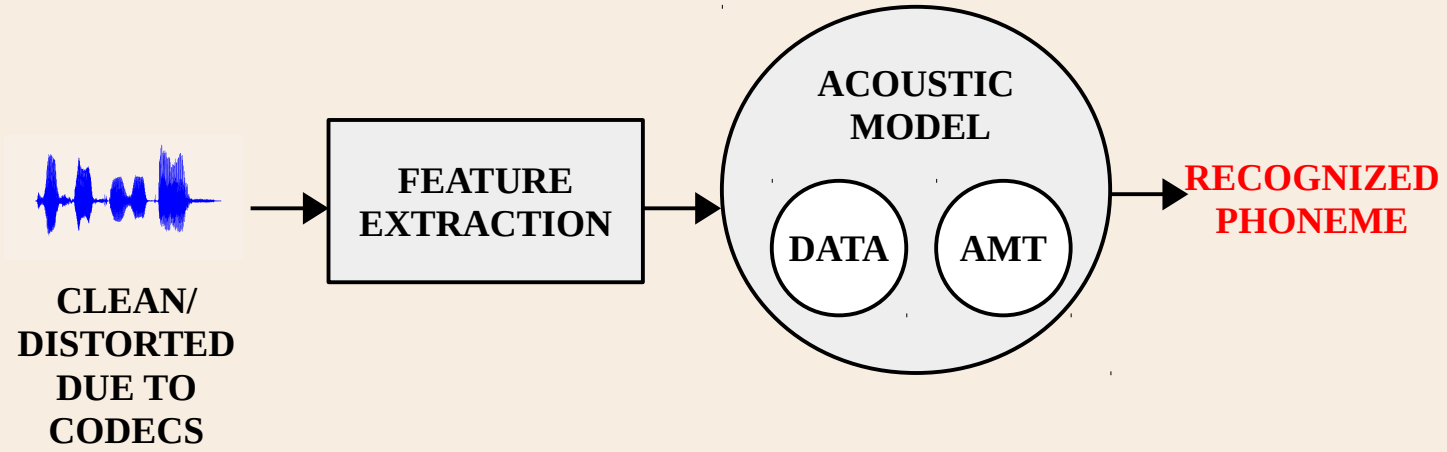


Single Encoding-Decoding

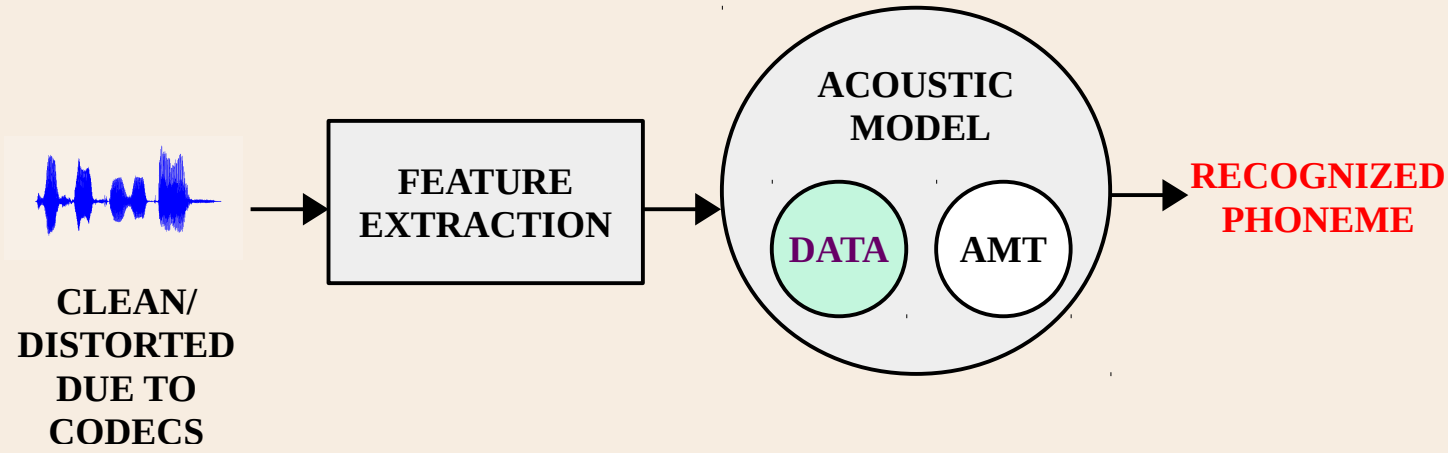


Tandem Encoding-Decoding

Problem statement



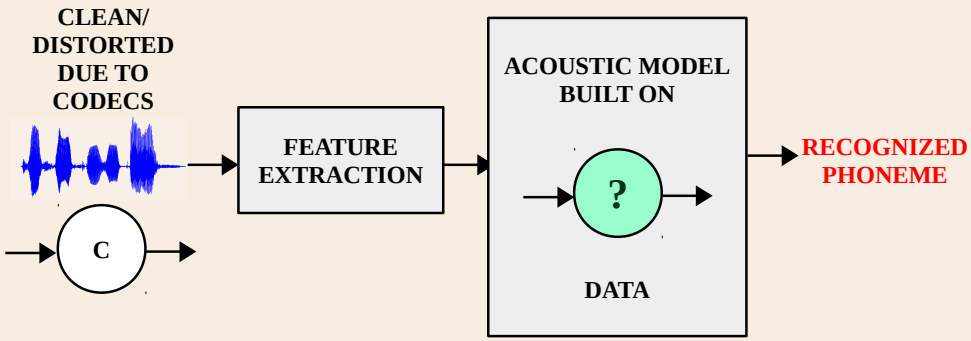
Problem statement



Problem statement

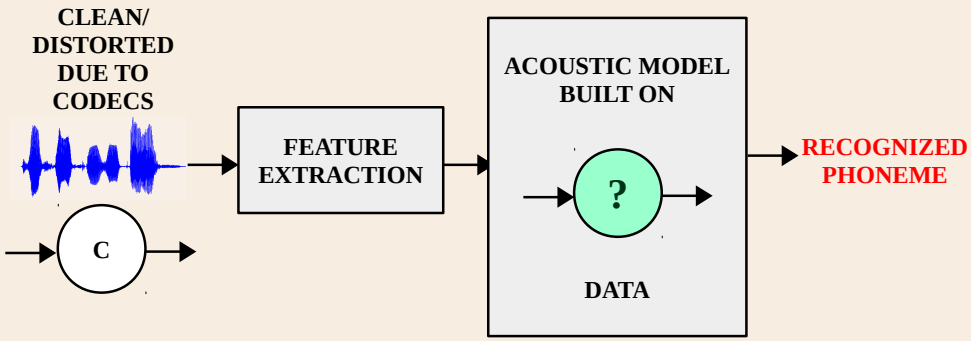
What is that **specific codec trained acoustic model**, that performs well for different types of input speech (coded or clean PCM) across different AMTs? **Robust to codec induced distortions.**

Problem statement

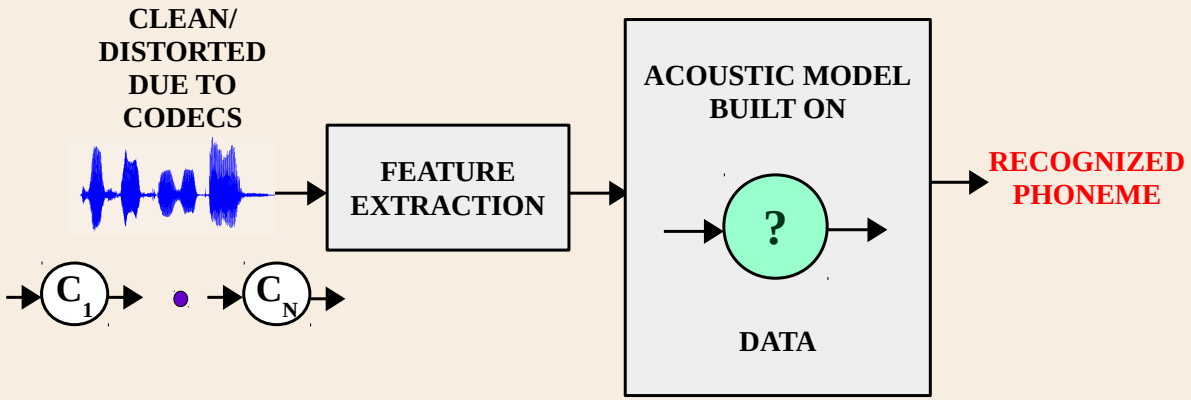


Single Encoding-Decoding

Problem statement

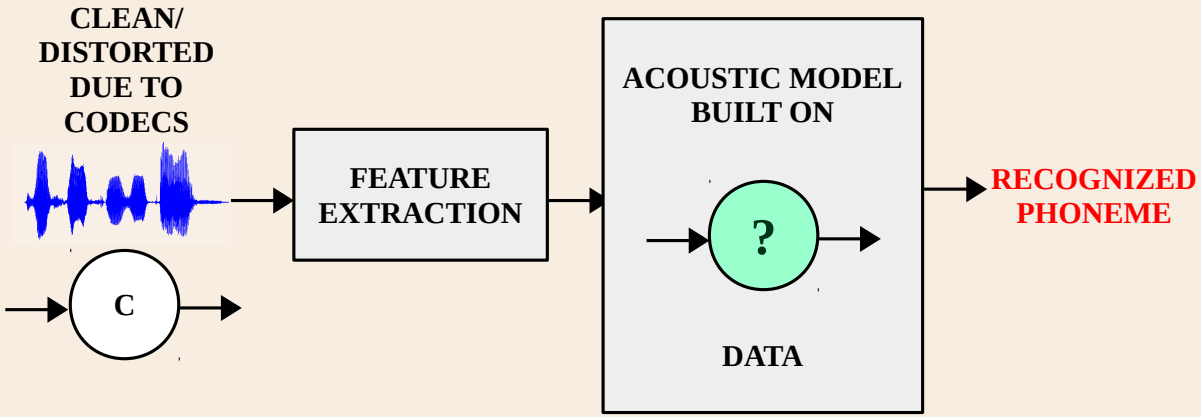


Single Encoding-Decoding

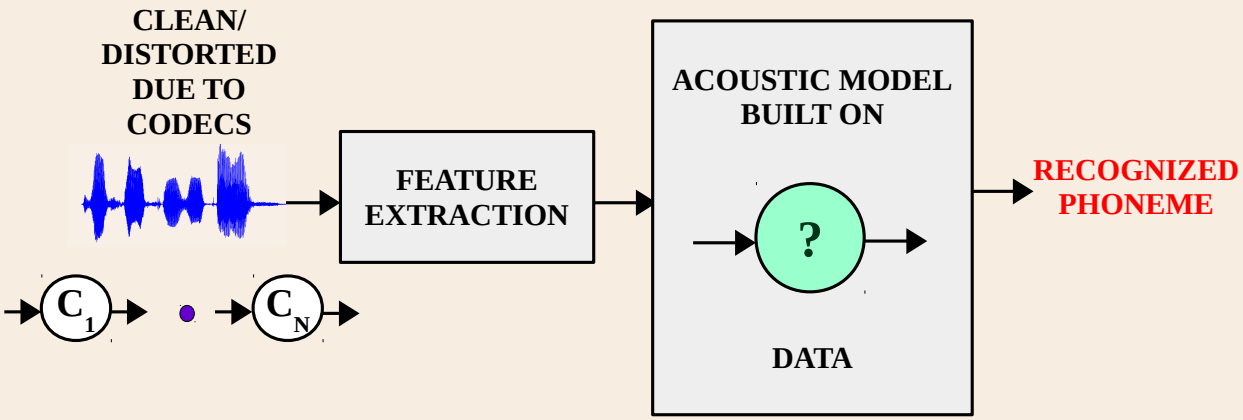


Tandem Encoding-Decoding

Key Finding 1

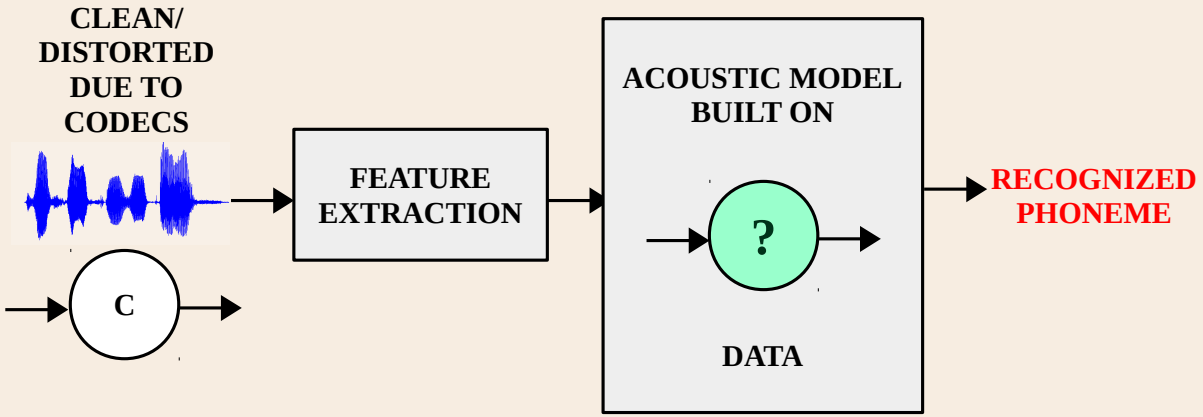


Single Encoding-Decoding

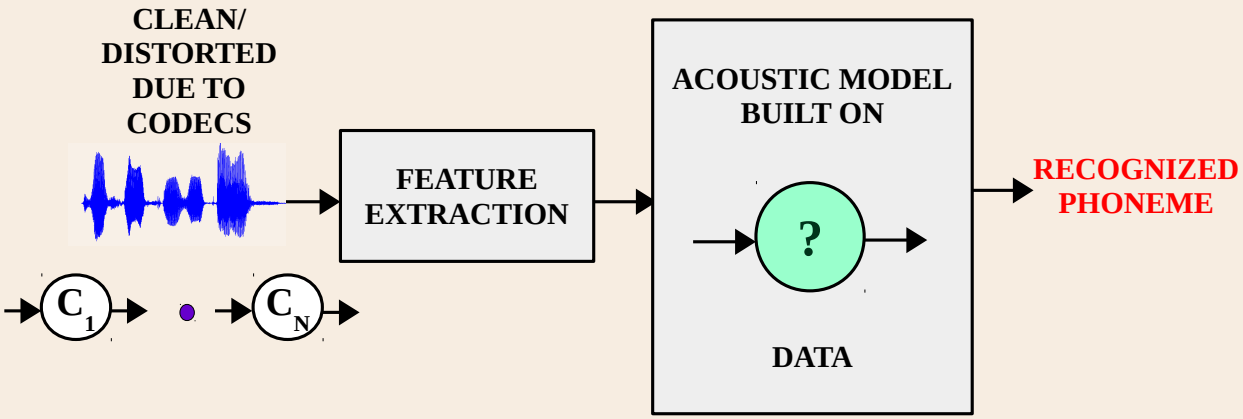


Tandem Encoding-Decoding

Key Finding 1

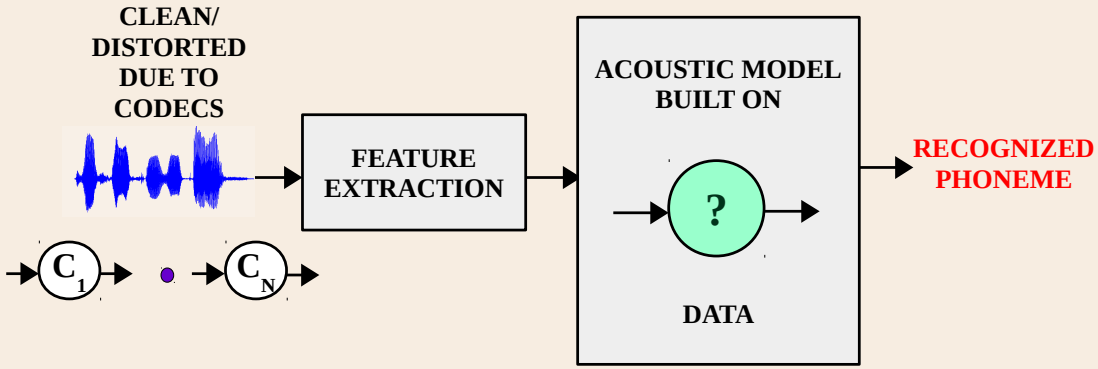


Single Encoding-Decoding



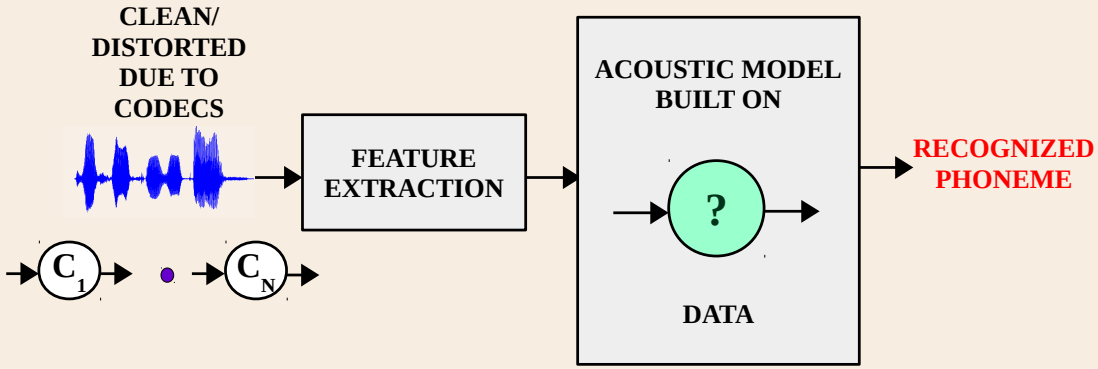
Tandem Encoding-Decoding

Key Finding 2

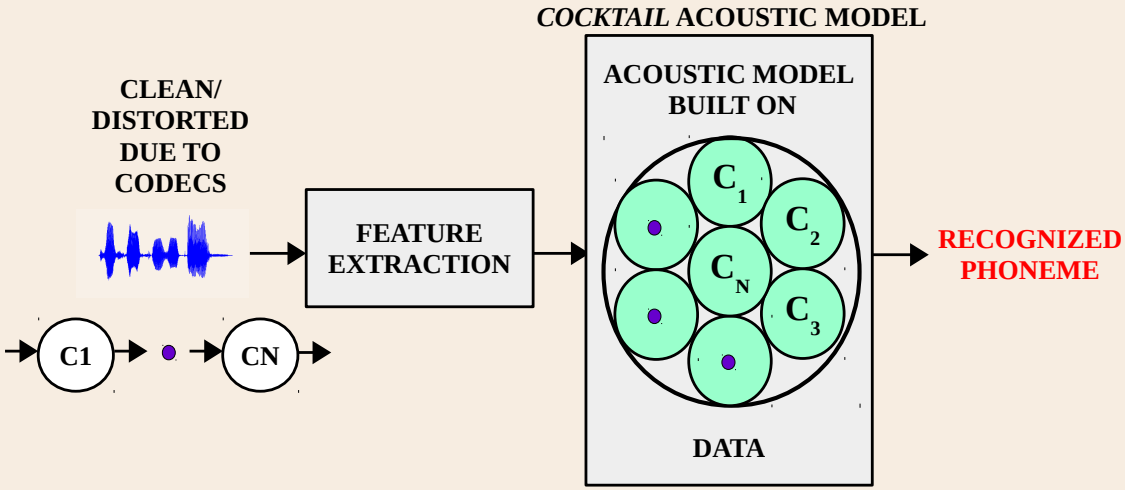


Tandem Encoding-Decoding

Key Finding 2



Tandem Encoding-Decoding



Cocktail Acoustic Model

Section 2

Existing Literature

Single Encoding-Decoding

1. Lower recognition for **low bit-rate codecs** [Euler *et al.* (1994), Lilly *et al.* (1996)].
2. Study of speech recognition with **GSM codecs** [Kim *et al.* (2000), H.-G. Hirsch (2002)].
3. ASR under **noisy conditions** using G.729, G.723.1 and GSM codecs [Grande *et al.* (2001)]

Tandem Encoding-Decoding

1. Impact on ASR performance more for **low bit-rate codecs** [Lilly *et al.* (1996)].
2. Study of ASR performance under **unkown Tandem scenario** [Salonidis *et al.* (1998)].

Compensation Strategies

1. **Enhancement** of the decoded speech, robust feature extraction [Dufour *et al.* (1996)]
2. **Adaptation** of acoustic models [Mokbel *et al.*. (1997), Salonidis *et al.* (1998), Srinivasamurthy *et al.* (2001)]

Section 3

AMTs and Codecs

Acoustic Modeling Techniques (AMT)

1. Monophone based GMM-HMM (MONO)
2. Context-dependent triphone based GMM-HMM (CD-TRI)
3. The Subspace Gaussian models with boosted Maximum Mutual Information (SGMM)
4. DNN with DBN Pretraining (DNN-DP)
5. DNN with state-level MBR (DNN-DP-sMBR)

Details

1. Kaldi toolkit [Povey *et al.* (2011)].
2. ASR metric: Phoneme Error Rate (PER)
3. Codecs source: IT-UT standards, SoX, SPEEX.
4. 0-gram language model.

List of codecs

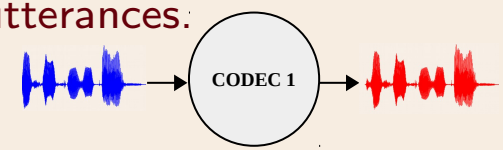
Codec	**Type**	**Band-width**	**Bit-Rate (kbps)**
G.711A	Waveform	Narrow	64
MELP	Parametric	Narrow	2.4
AMR-NB	Hybrid	Narrow	4.40
AMR-WB	Hybrid	Wide	23.85
G.728	Hybrid	Narrow	16
G.729A	Hybrid	Narrow	8
G.729B	Hybrid	Narrow	8
PCM	Waveform	Narrow	128
ADPCM	Waveform	Wide	32
GSM-8k	Hybrid	Narrow	13
SPEEX	Hybrid	Wide	27.8

Datasets

- TIMIT database. Sampling rate: 8kHz.
- Training set: 462 speakers with 3696 utterances.
- Development Set: 50 speakers with 400 utterances.
- Test Set: 24 speakers with 192 utterances.

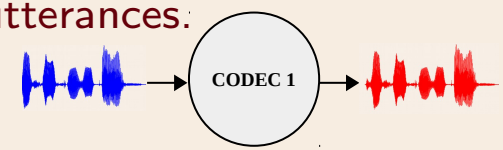
Datasets

- TIMIT database. Sampling rate: 8kHz.
- Training set: 462 speakers with 3696 utterances.
- Development Set: 50 speakers with 400 utterances.
- Test Set: 24 speakers with 192 utterances.



Datasets

- TIMIT database. Sampling rate: 8kHz.
- Training set: 462 speakers with 3696 utterances.
- Development Set: 50 speakers with 400 utterances.
- Test Set: 24 speakers with 192 utterances.

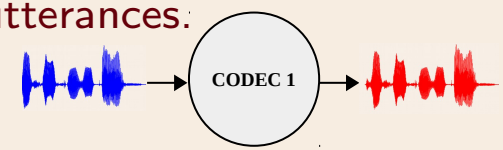


- 8 acoustic models using single encoding-decoding.

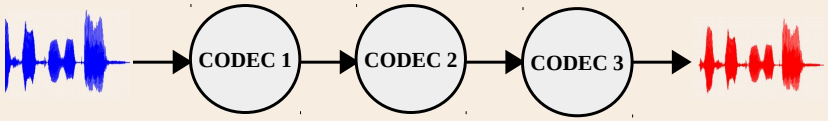
Codec	Type	Band-width	Bit-rate (kbps)
G.711A	Waveform	Narrow	64
MELP	Parametric	Narrow	2.4
AMR-NB	Hybrid	Narrow	4.40
AMR-WB	Hybrid	Wide	23.85
G.728	Hybrid	Narrow	16
G.729A	Hybrid	Narrow	8
G.729B	Hybrid	Narrow	8
PCM	Waveform	Narrow	128
ADPCM	Waveform	Wide	32
GSM-8k	Hybrid	Narrow	13
SPEEX	Hybrid	Wide	27.8

Datasets

- TIMIT database. Sampling rate: 8kHz.
- Training set: 462 speakers with 3696 utterances.
- Development Set: 50 speakers with 400 utterances.
- Test Set: 24 speakers with 192 utterances.



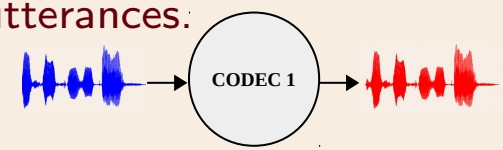
- 8 acoustic models using single encoding-decoding.



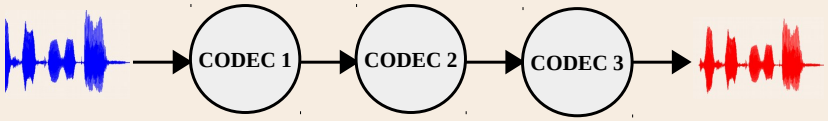
Codec	Type	Band-width	Bit-rate (kbps)
G.711A	Waveform	Narrow	64
MELP	Parametric	Narrow	2.4
AMR-NB	Hybrid	Narrow	4.40
AMR-WB	Hybrid	Wide	23.85
G.728	Hybrid	Narrow	16
G.729A	Hybrid	Narrow	8
G.729B	Hybrid	Narrow	8
PCM	Waveform	Narrow	128
ADPCM	Waveform	Wide	32
GSM-8k	Hybrid	Narrow	13
SPEEX	Hybrid	Wide	27.8

Datasets

- TIMIT database. Sampling rate: 8kHz.
- Training set: 462 speakers with 3696 utterances.
- Development Set: 50 speakers with 400 utterances.
- Test Set: 24 speakers with 192 utterances.



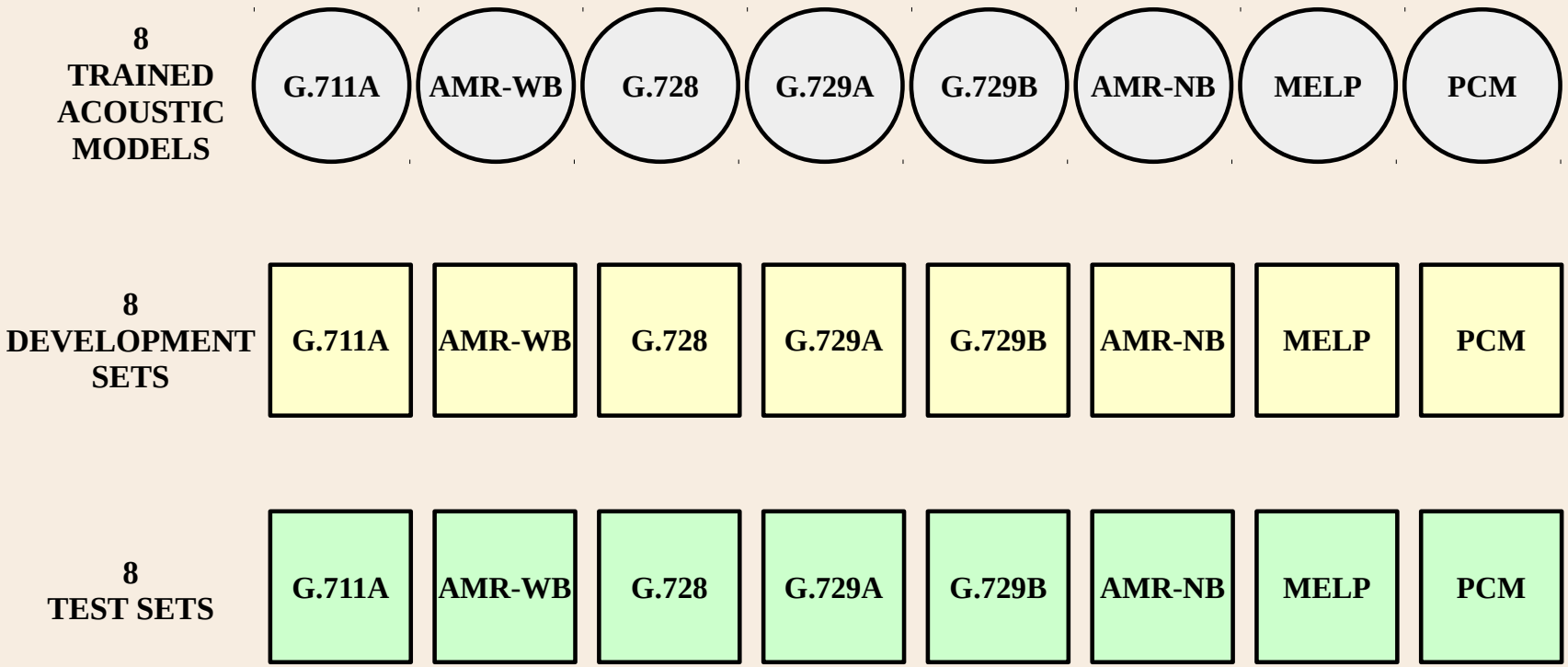
- 8 acoustic models using single encoding-decoding.



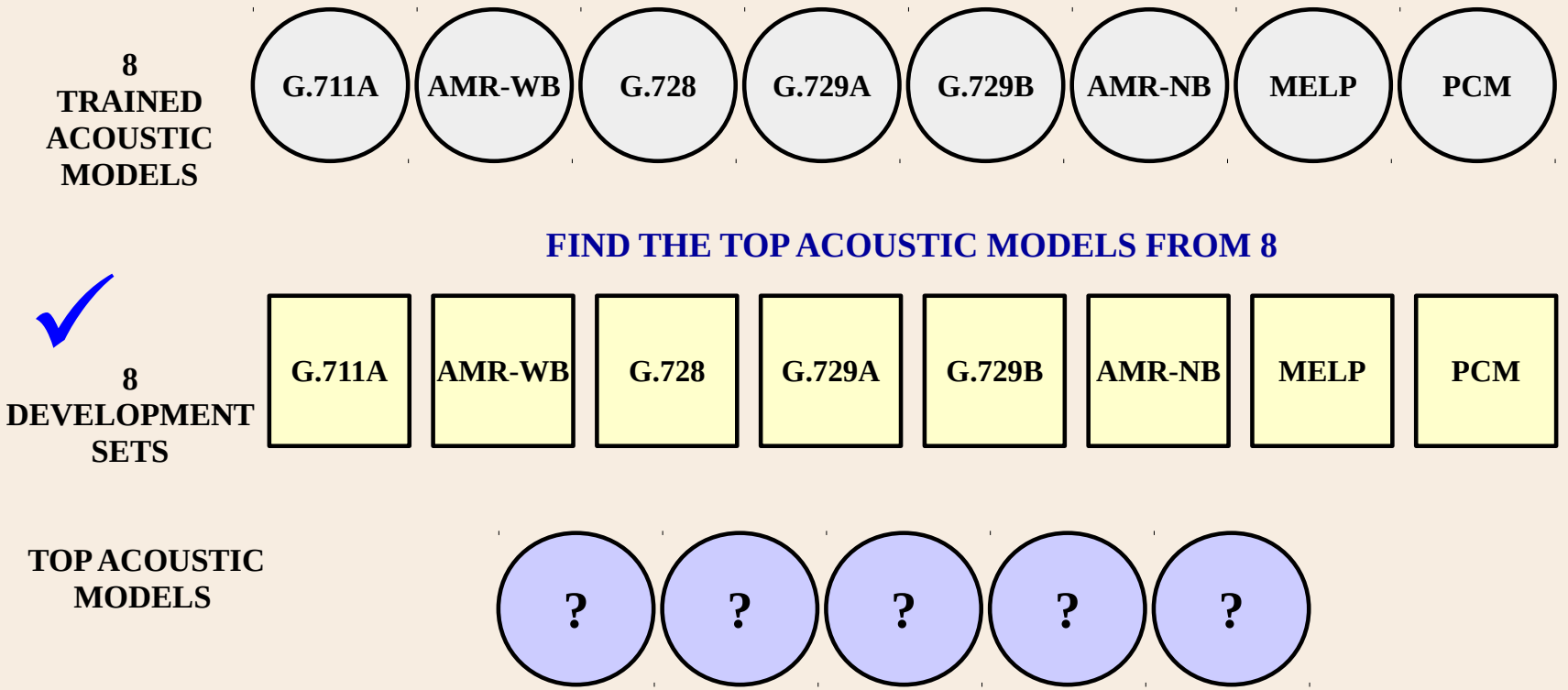
Codec	Type	Band-width	Bit-rate (kbps)
G.711A	Waveform	Narrow	64
MELP	Parametric	Narrow	2.4
AMR-NB	Hybrid	Narrow	4.40
AMR-WB	Hybrid	Wide	23.85
G.728	Hybrid	Narrow	16
G.729A	Hybrid	Narrow	8
G.729B	Hybrid	Narrow	8
PCM	Waveform	Narrow	128
ADPCM	Waveform	Wide	32
GSM-8k	Hybrid	Narrow	13
SPEEX	Hybrid	Wide	27.8

- 6 Tandem test databases: 1) ADPCM→GSM-8k→SPEEX, 2) ADPCM→SPEEX→GSM-8k, 3) GSM-8k→ADPCM→SPEEX, 4) GSM-8k→SPEEX→ADPCM, 5) SPEEX→ADPCM→GSM-8k, 6) SPEEX→GSM-8k→ADPCM.

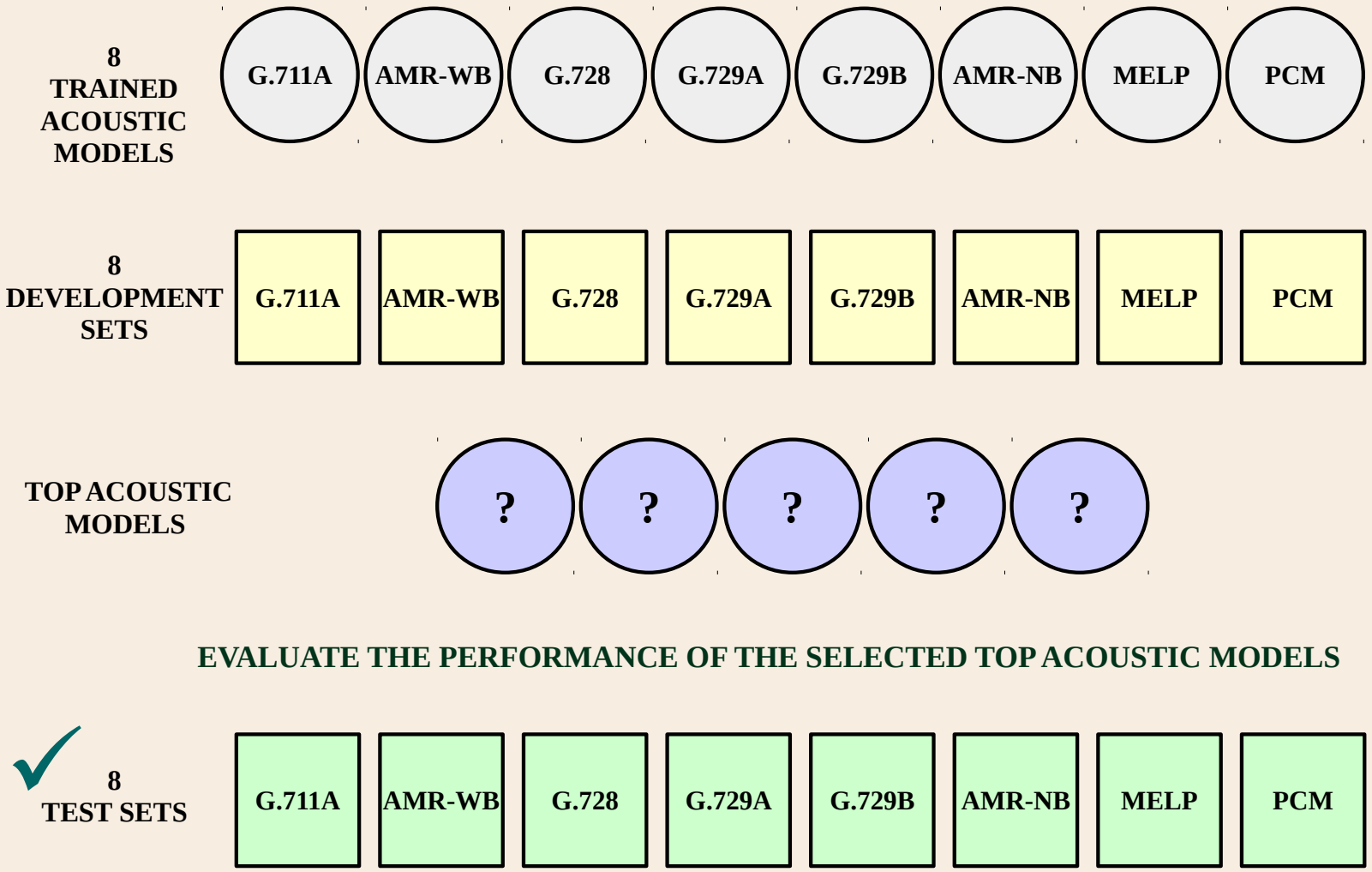
Overview of Experiments: Single Encoding Decoding



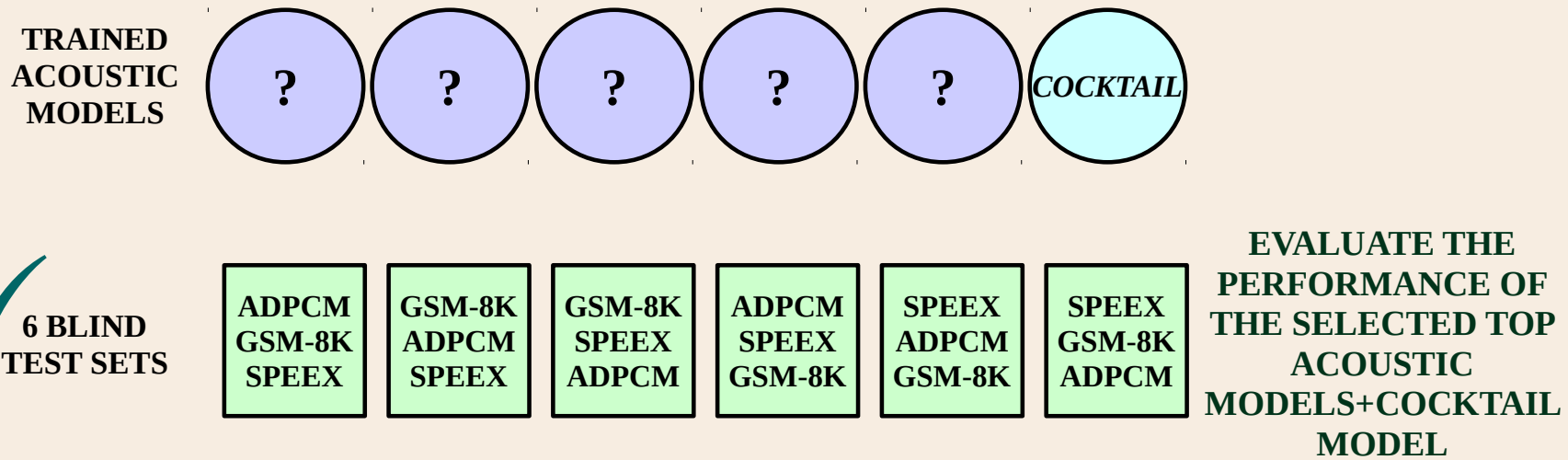
Overview of Experiments: Single Encoding Decoding



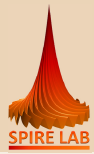
Overview of Experiments: Single Encoding Decoding



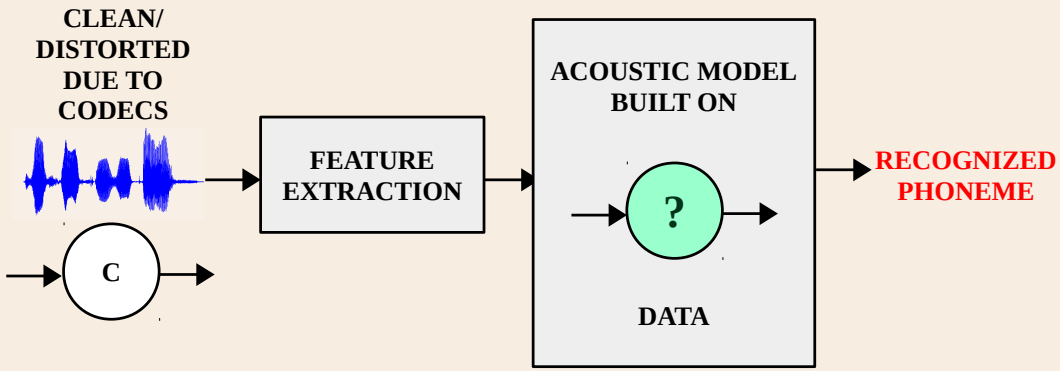
Overview of Experiments: Tandem Encoding Decoding



Section 4



Single Encoding Decoding



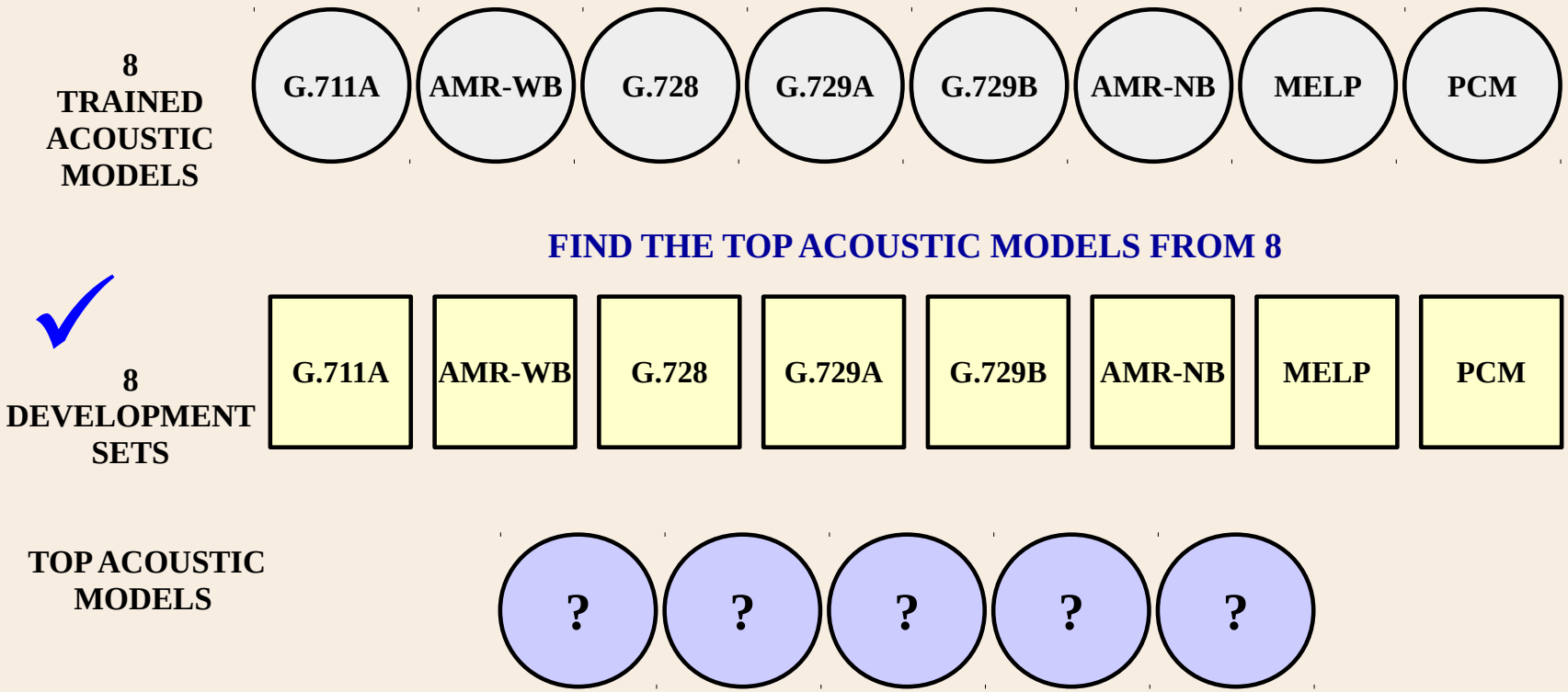
Single Encoding-Decoding

Question

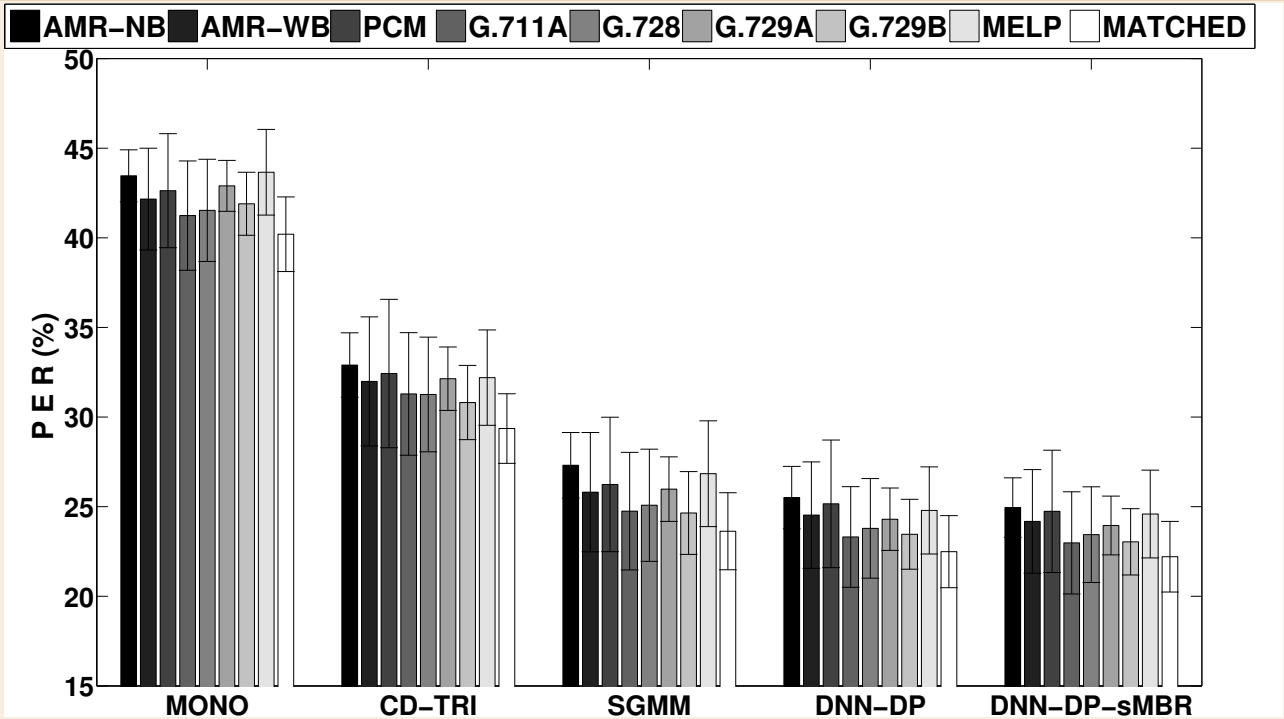
What are the **best acoustic models** across all the AMTs for various coded speech?

- 8 Candidate Models: G.711A, MELP, AMR-NB, AMR-WB, G.728, G.729A, G.729B, PCM.
- 8 development and 8 test datasets.

Single Encoding Decoding: Choice of Top codecs

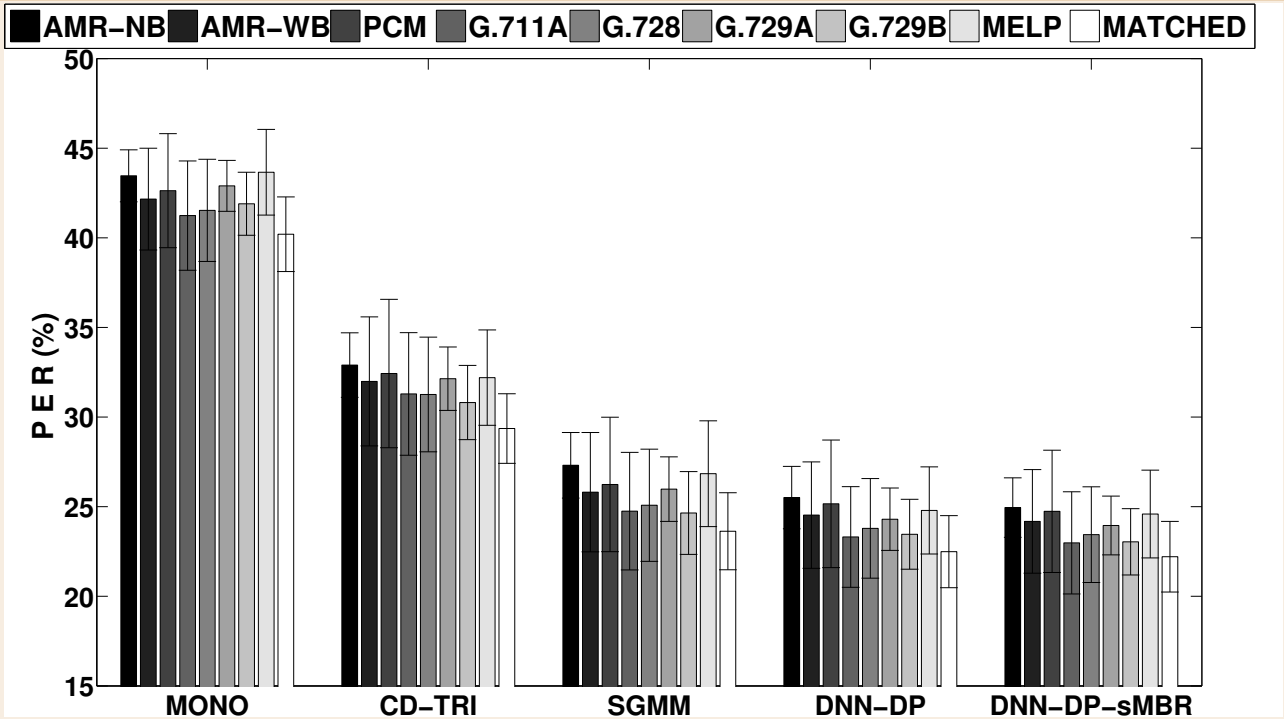


Single Encoding Decoding: Choice of Top codecs



The average (standard deviation) PER (%) for 8 acoustic models and 5 AMTs across the **development sets**.

Single Encoding Decoding: Choice of Top codecs

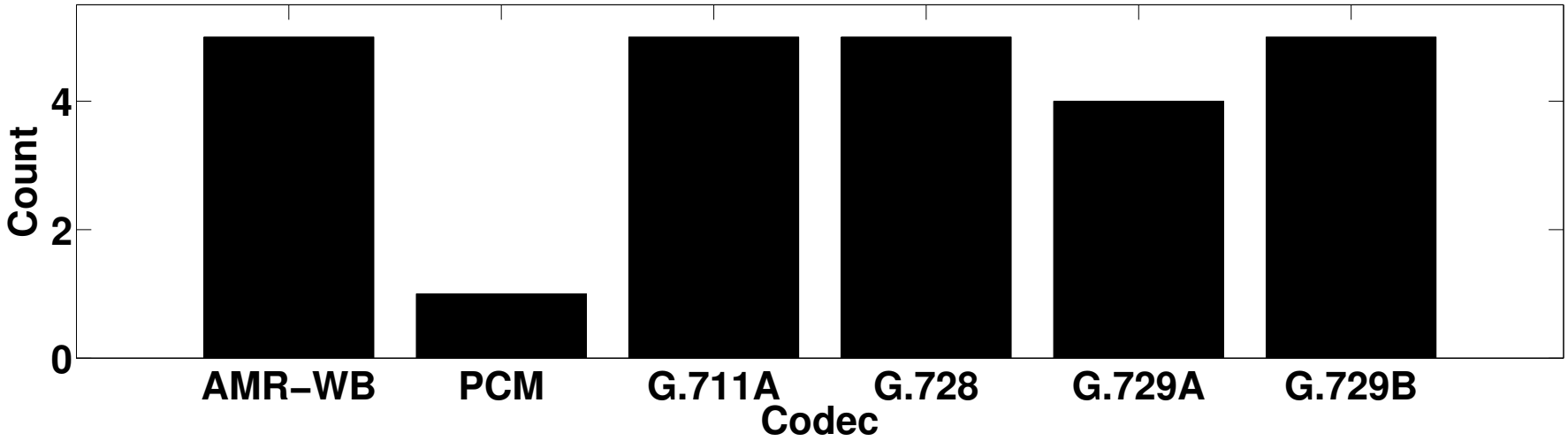


The average (standard deviation) PER (%) for 8 acoustic models and 5 AMTs across the **development sets**.

Results

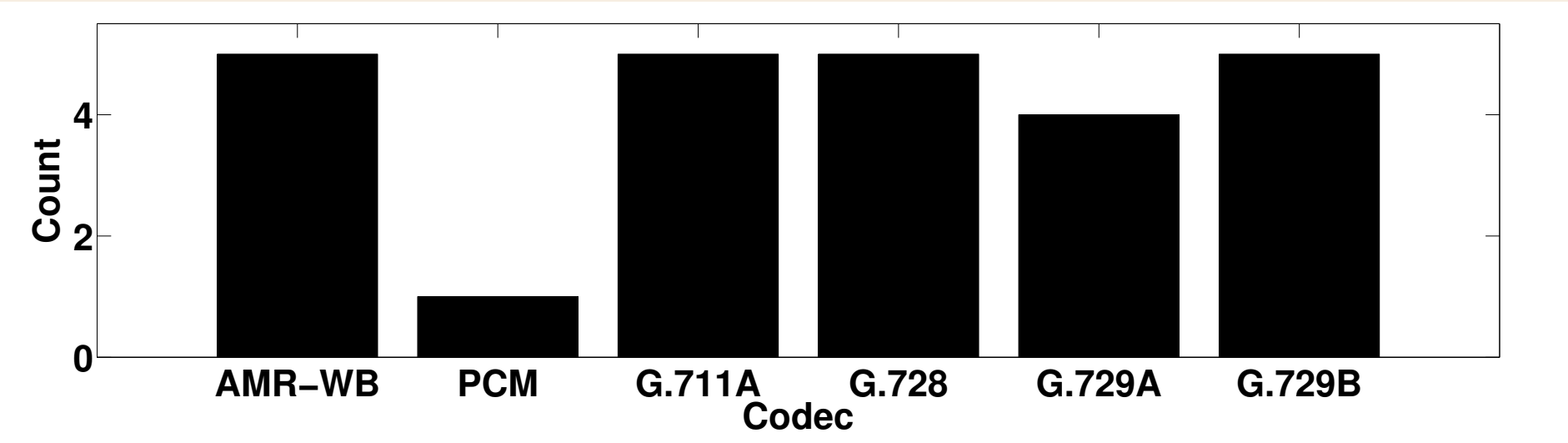
- PER decreases with the improvements in the AMTs.
- Matched condition performs best across all the AMTs.

Single Encoding Decoding: Choice of Top codecs



Histogram of top four ranked codecs across different AMTs.

Single Encoding Decoding: Choice of Top codecs

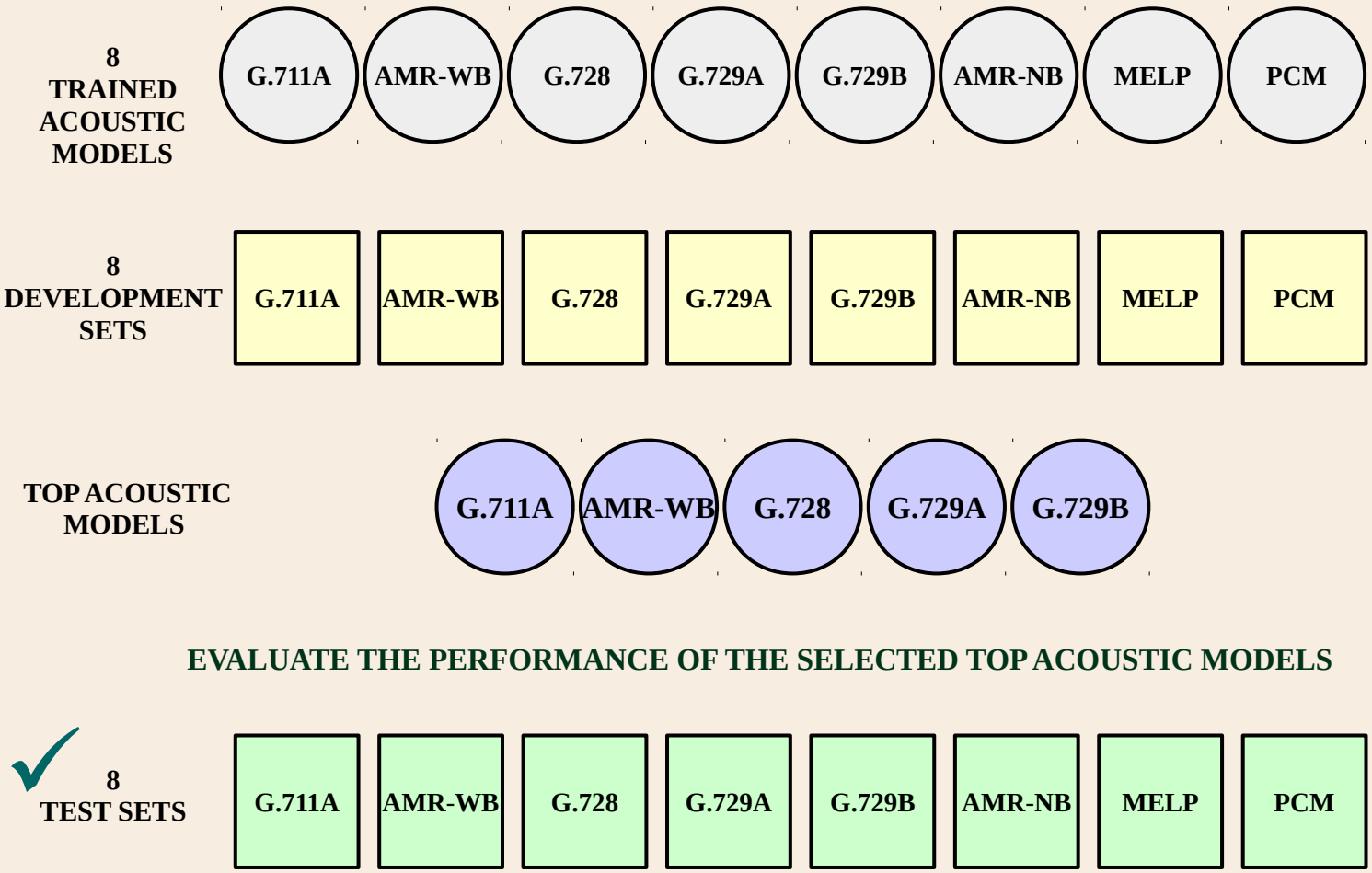


Histogram of top four ranked codecs across different AMTs.

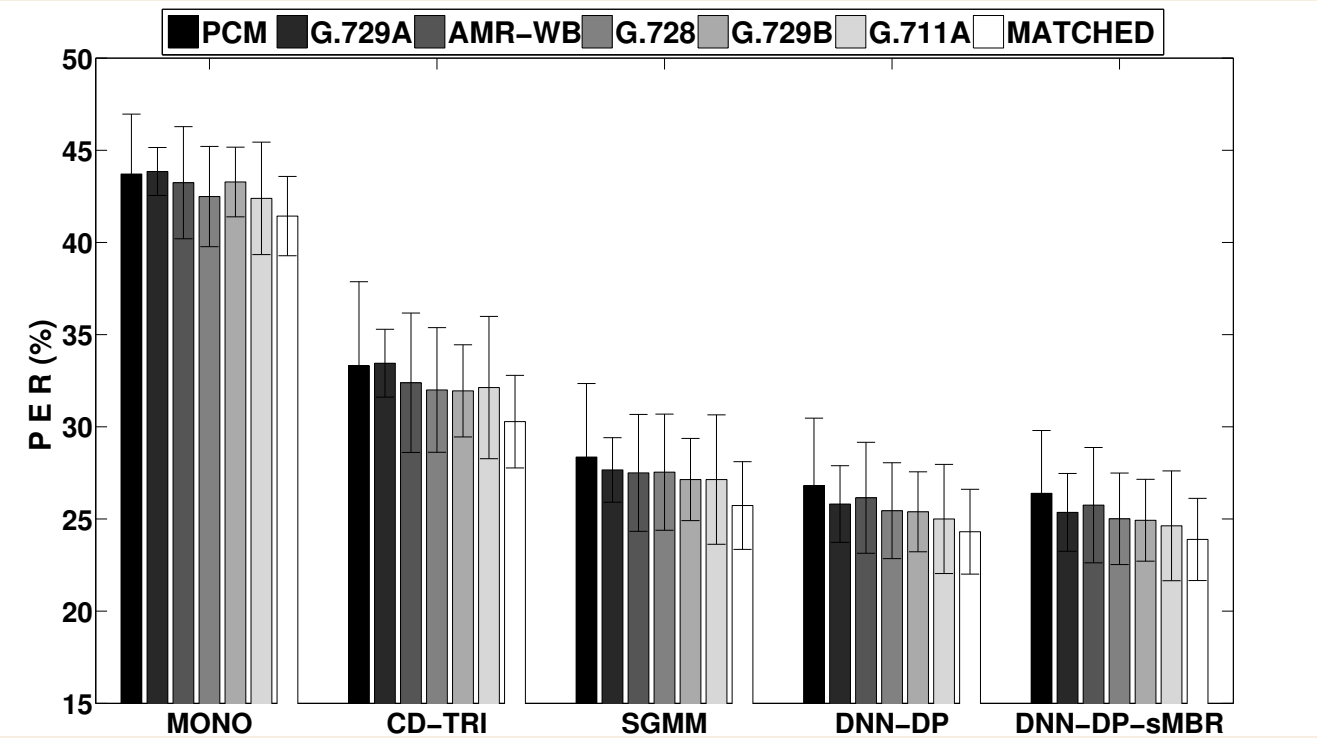
Results

- Higher bit rate codecs.
- Most of them are narrowband codecs.

Single Encoding Decoding: Performance of top codecs

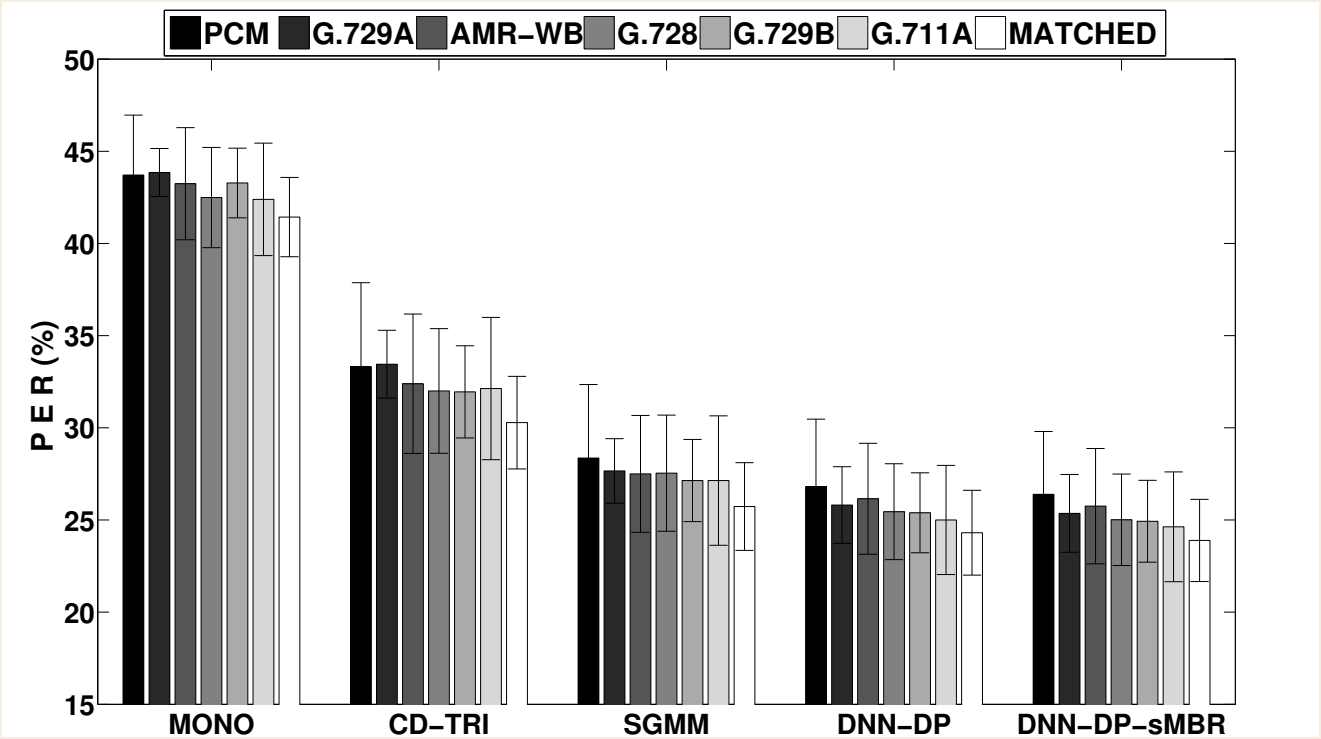


Single Encoding Decoding: Performance of top codecs



The average (standard deviation) PER (%) for the top 5 acoustic models (along with PCM and Mixed) and 5 AMTs across the **test sets**

Single Encoding Decoding: Performance of top codecs

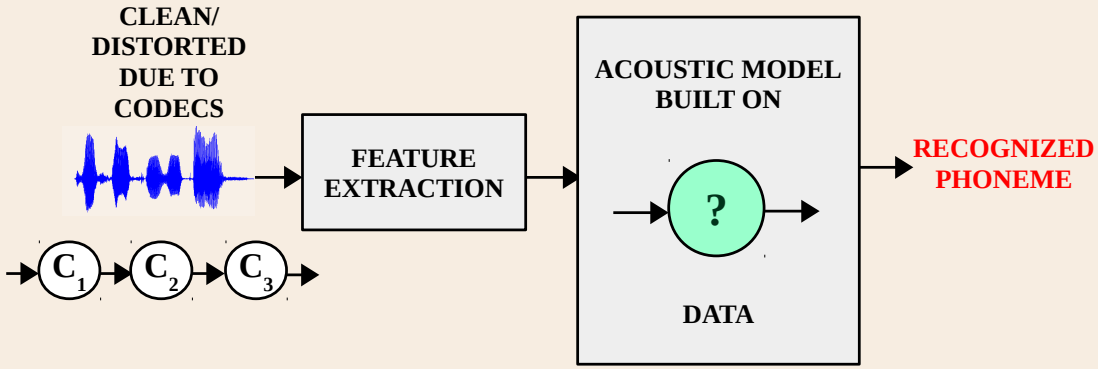


The average (standard deviation) PER (%) for the top 5 acoustic models (along with PCM and Mixed) and 5 AMTs across the **test sets**

Results

- PER decreases with the improvements in the AMTs.
- Least PER for **G.711A** based acoustic model.

Tandem Encoding Decoding



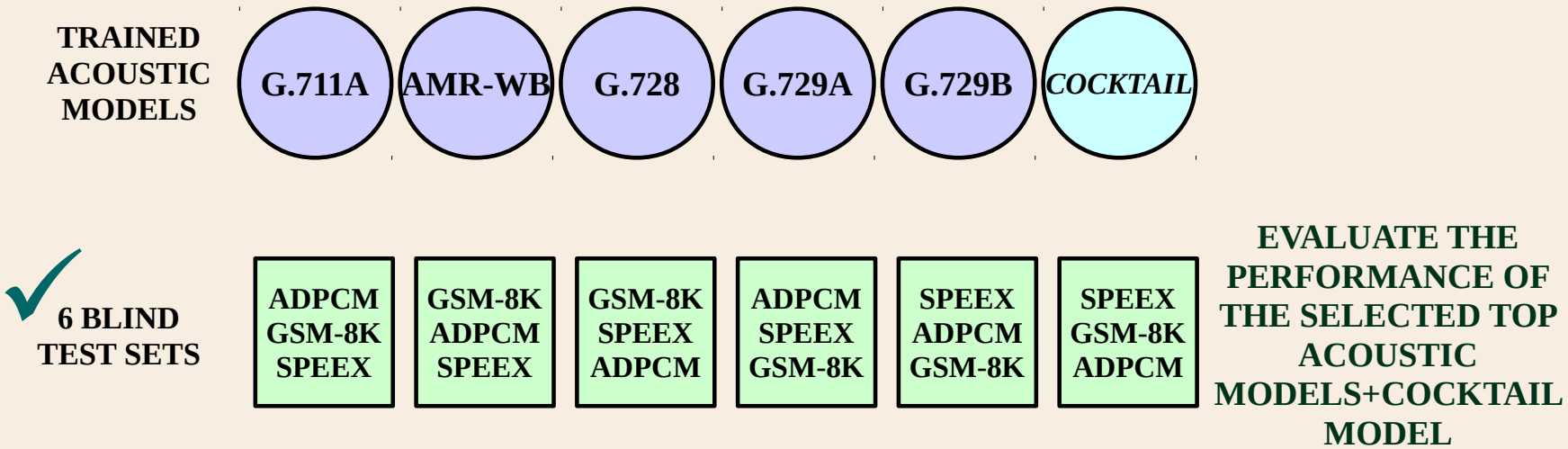
Tandem Encoding-Decoding

Question

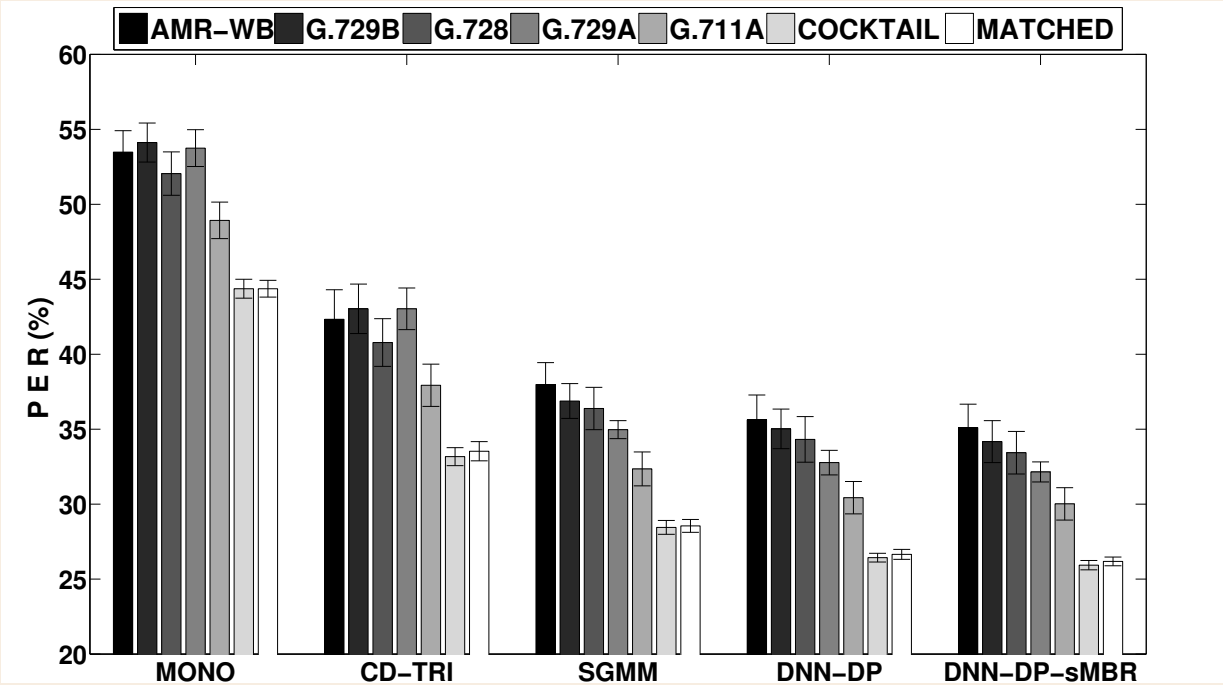
How do the **top five acoustic models** perform across all the AMTs for tandem coded speech?

- 6 Candidate models: G.711A, AMR-WB, G.728, G.729A, G.729B, Cocktail.
- 6 blind test sets: Combinations of ADPCM, GSM-8k, SPEEX.

Tandem Encoding Decoding: Performance of top codecs

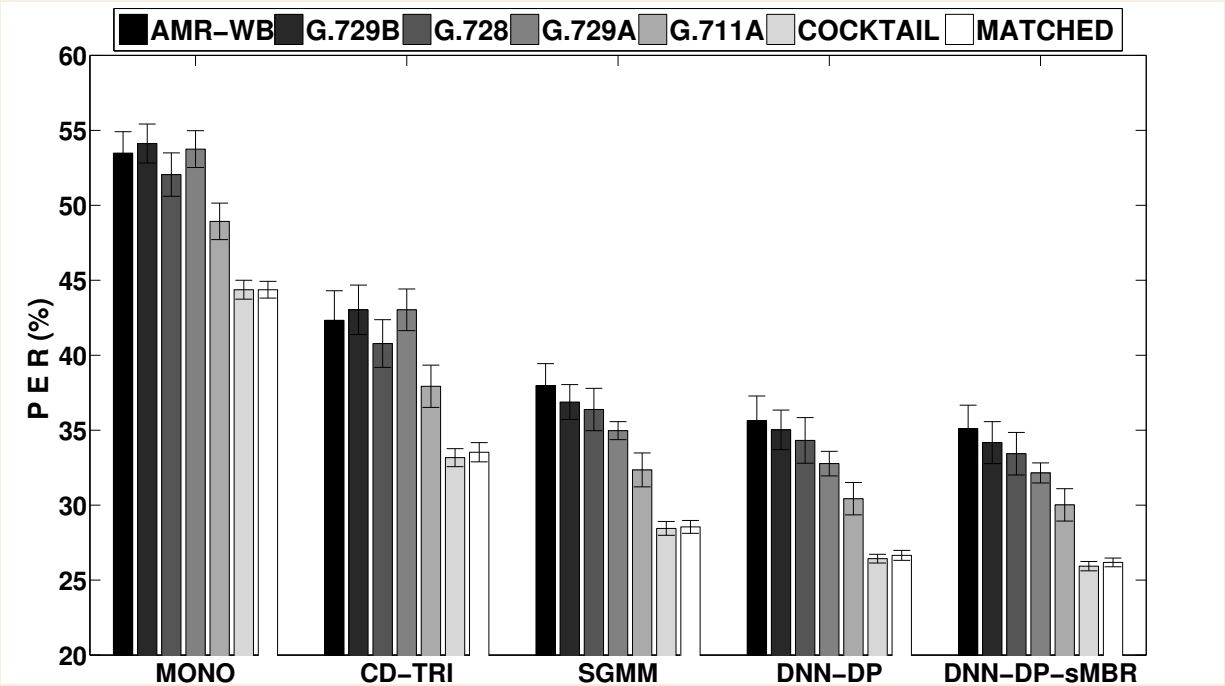


Tandem Encoding Decoding: Performance of top codecs



The average (standard deviation) PER (%) for 6 acoustic models and 5 AMTs across six **blind test sets**

Tandem Encoding Decoding: Performance of top codecs



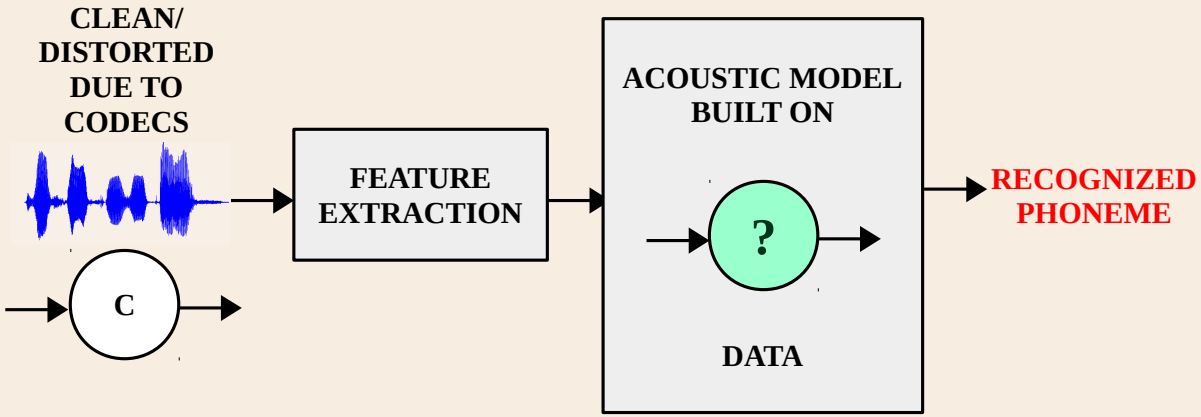
The average (standard deviation) PER (%) for 6 acoustic models and 5 AMTs across six **blind test sets**

Results

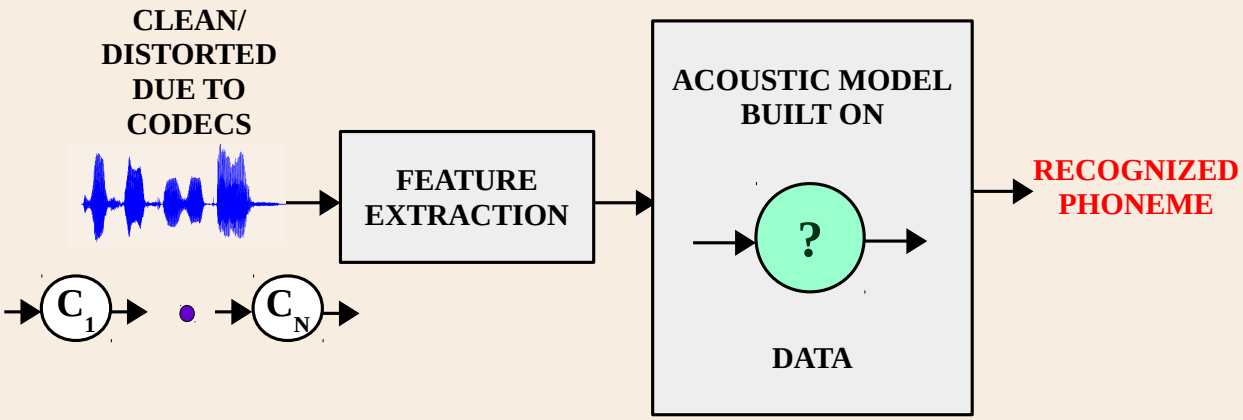
- PER decreases with the improvements in the AMTs.
- Least PER for **G.711A** based acoustic model.
- *Cocktail* acoustic model is comparable to the matched condition.

Section 5

Key Finding 1

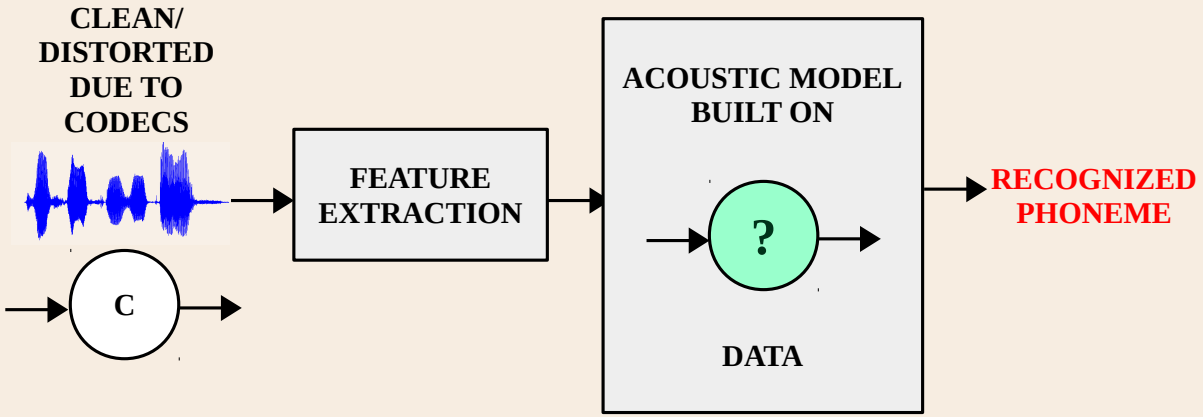


Single Encoding-Decoding

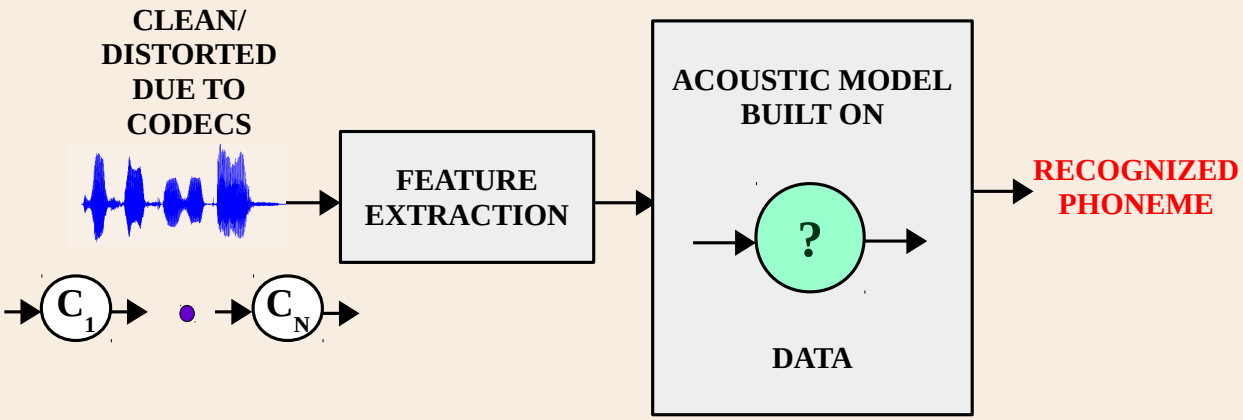


Tandem Encoding-Decoding

Key Finding 1



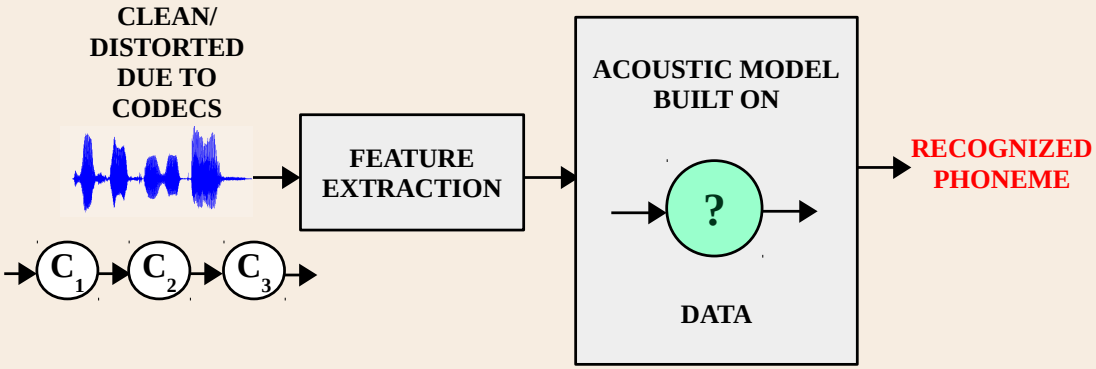
Single Encoding-Decoding



Tandem Encoding-Decoding

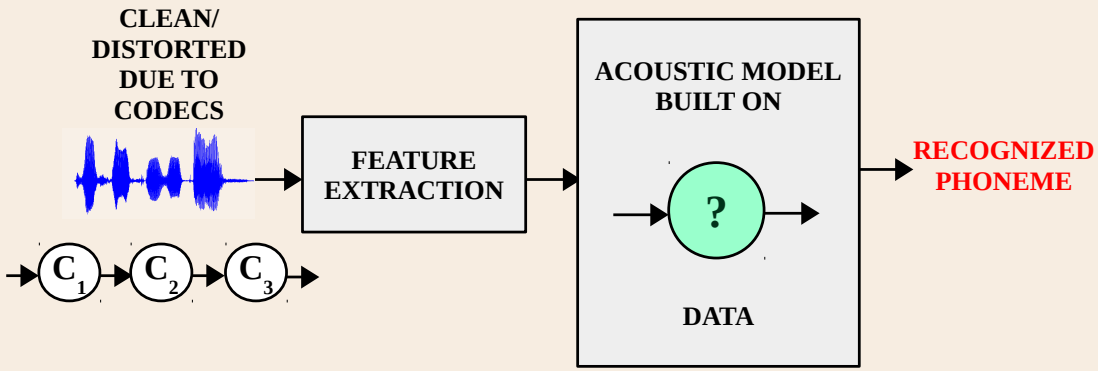
Narrowband High bit-rate codec

Key Finding 2

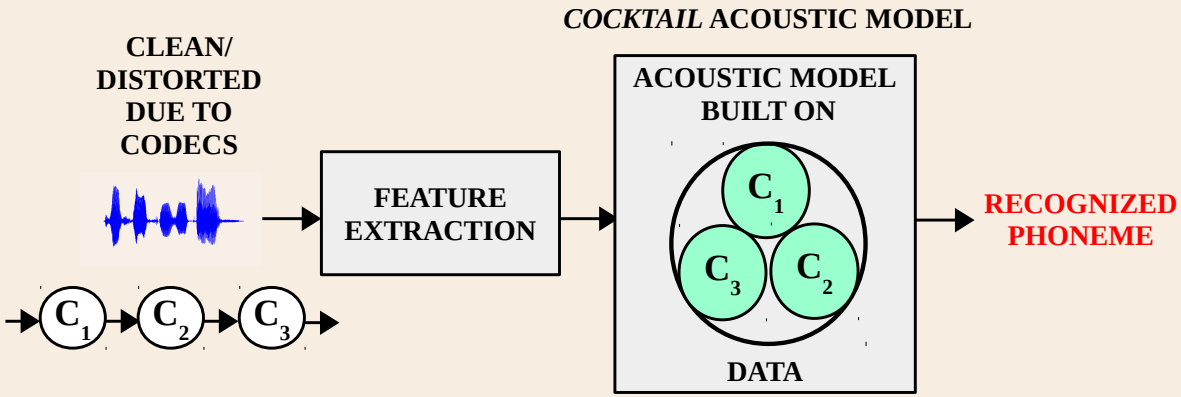


Tandem Encoding-Decoding

Key Finding 2

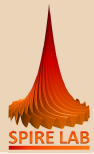


Tandem Encoding-Decoding



Cocktail Acoustic Model

Summary



Conclusions

1. Studied the **codec induced distortion** on the ASR performance.

Summary

Conclusions

1. Studied the **codec induced distortion** on the ASR performance.
2. **G.711A, a narrowband high bit rate codec**, results in the best ASR accuracy.

Summary

Conclusions

1. Studied the **codec induced distortion** on the ASR performance.
2. **G.711A, a narrowband high bit rate codec**, results in the best ASR accuracy.
3. If the pool of **tandem topologies are known a priori**, cocktail acoustic model could be used.

Summary

Conclusions

1. Studied the **codec induced distortion** on the ASR performance.
2. **G.711A, a narrowband high bit rate codec**, results in the best ASR accuracy.
3. If the pool of **tandem topologies are known a priori**, cocktail acoustic model could be used.

Future works

1. Effectiveness of the best performing models along with **language models**.

Summary

Conclusions

1. Studied the **codec induced distortion** on the ASR performance.
2. **G.711A, a narrowband high bit rate codec**, results in the best ASR accuracy.
3. If the pool of **tandem topologies are known a priori**, cocktail acoustic model could be used.

Future works

1. Effectiveness of the best performing models along with **language models**.
2. **Compensation** of the codec induced distortions to aid ASR.

THANK YOU