



# Phonetic Error Analysis Beyond Phone Error Rate

Erfan Loweimi

CSTR Talk, University of Edinburgh  
11, Dec, 2023

## We will look at ...

3346

IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 31, 2023

# Phonetic Error Analysis Beyond Phone Error Rate

Erfan Loweimi , Member, IEEE, Andrea Carmantini , Member, IEEE, Peter Bell , Steve Renals , Fellow, IEEE, and Zoran Cvetkovic , Senior Member, IEEE

E. Loweimi, A. Carmantini, P. Bell, S. Renals and Z. Cvetkovic, “*Phonetic Error Analysis Beyond Phone Error Rate*”, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3346-3361, 2023, doi: 10.1109/TASLP.2023.3313417.

# Outline

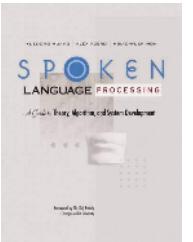
- Analysis beyond PER
  - What / How / Why
- Effect of Various Factors

# Analysis beyond PER ...

- What is the contribution of each broad phonetic class (BPC)  $c$  in PER?

$$PER = \sum_c (PER_c) ?$$

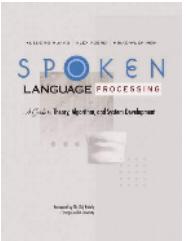
# Three BPCs considered ...



classes	phones
Affricates	ch jh
Diphthongs	aw ay ey ow oy
Fricatives	dh f s sh th v z
Nasal	m n ng
Plosive	b d dx g k p t
Semi-vowel	hh l r w y
Vowel	aa ae ah eh er ih iy uh uw
Silence	sil

(A) 8-class

# Three BPCs considered ...



classes	phones
Affricates	ch jh
Diphthongs	aw ay ey ow oy
Fricatives	dh f s sh th v z
Nasal	m n ng
Plosive	b d dx g k p t
Semi-vowel	hh l r w y
Vowel	aa ae ah eh er ih iy uh uw
Silence	sil

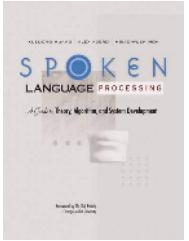
(A) 8-class

## (B) 3-class

classes	phones
Vowel <sup>+</sup>	aw ay ey ow oy aa ae ah eh er ih iy uh uw
Consonant	b ch d dh dx f g hh jh k l m n ng p r s sh t th v w y z
Silence	sil
Voiced	aa ae ah aw ay b d dh dx eh eer ey g hh ih iy jh l m n ng ow oy r uh uw v w y z
Unvoiced	ch f k p s sh t th

## (C) 3-class

# Three BPCs considered ...



classes	phones
Affricates	ch jh
Diphthongs	aw ay ey ow oy
Fricatives	dh f s sh th v z
Nasal	m n ng
Plosive	b d dx g k p t
Semi-vowel	hh l r w y
Vowel	aa ae ah eh er ih iy uh uw
Silence	sil

(A) 8-class

$$\text{Vowel}^+ = \text{Vowel} \cup \text{Diphthong}$$

$$\text{Silence} = /h\#/\cup/epi/\cup/pau/\cup\text{Closures}$$

Detail in (A1)



Loweimi et al

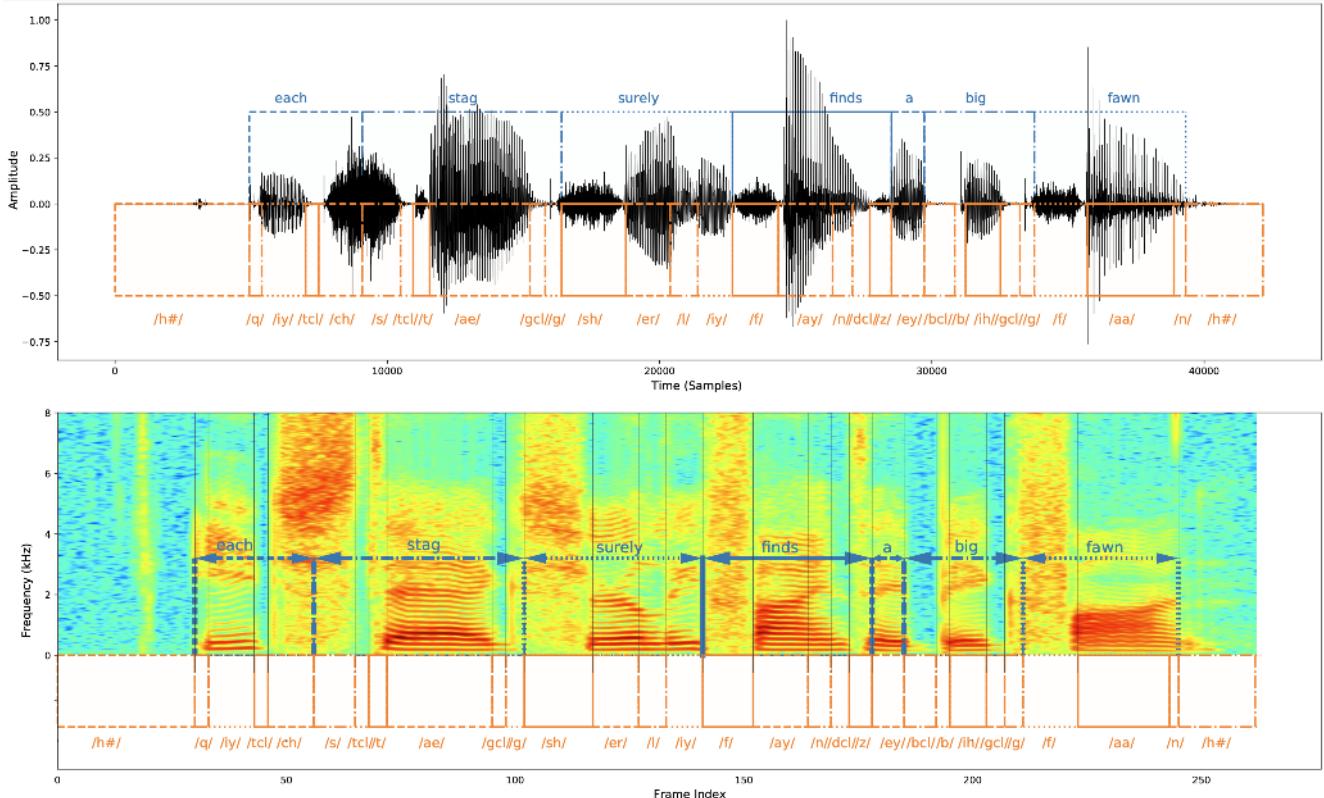
(B) 3-class

classes	phones
Vowel <sup>+</sup>	aw ay ey ow oy aa ae ah eh er ih iy uh uw
Consonant	b ch d dh dx f g hh jh k l m n ng p r s sh t th v w y z
Silence	sil
Voiced	aa ae ah aw ay b d dh dx eh eer ey g hh ih iy jh l m n ng ow oy r uh uw v w y z
Unvoiced	ch f k p s sh t th

(C) 3-class

# Phonetic Transcription

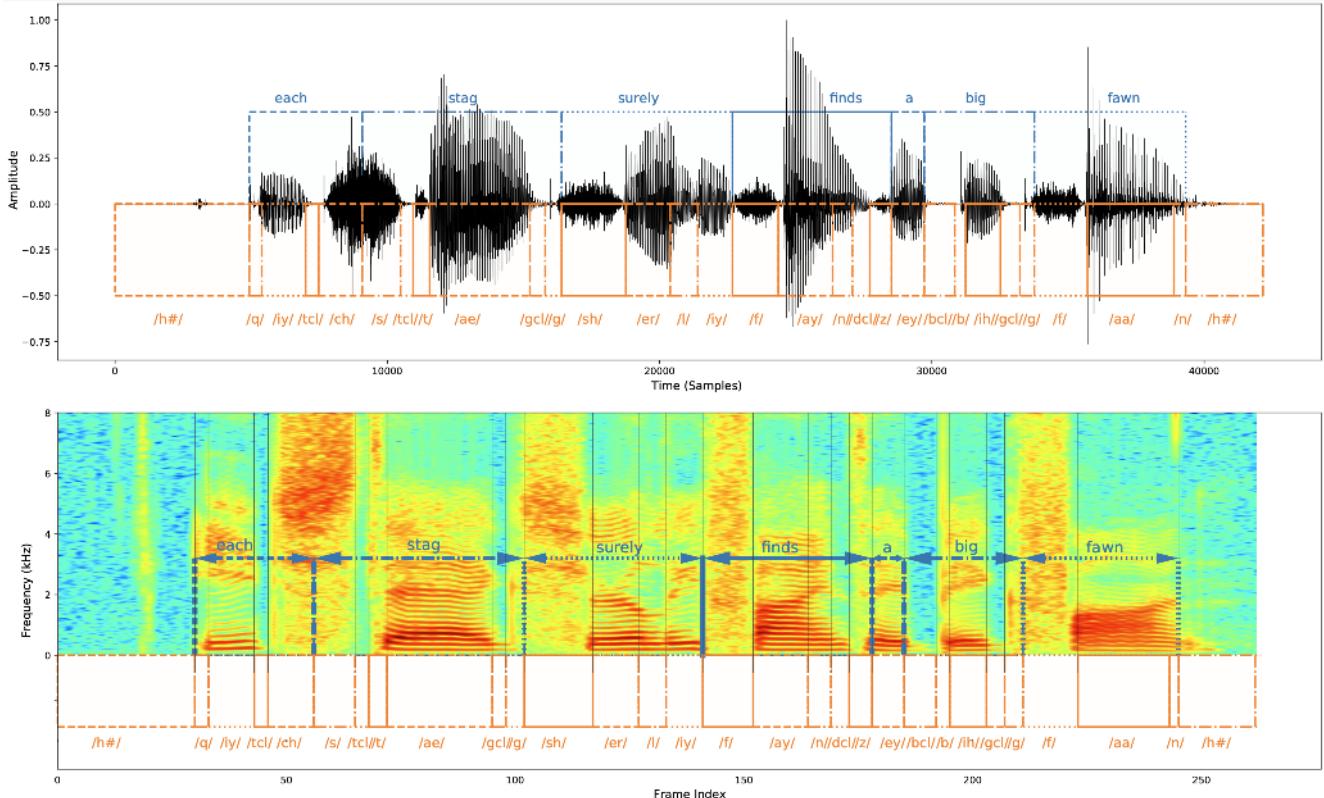
Data: TIMIT



- Original Transcription: **61** phones

# Phonetic Transcription

Data: TIMIT



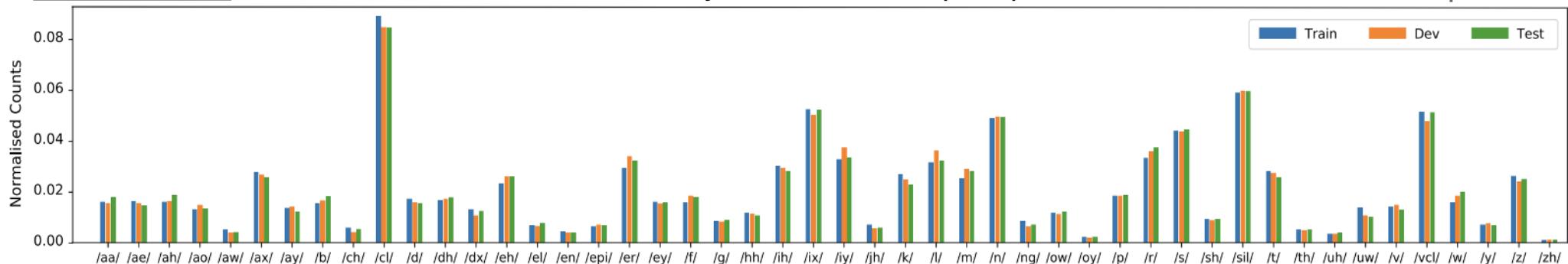
- Original Transcription: **61** phones → Train w/ **48** phones → Decode w/ **39** phones
- Mapping (Kaldi): `phones.60-48-39.map`

# Phonetic Distribution

Data: TIMIT

Probability Mass Function (PMF)

48 phones

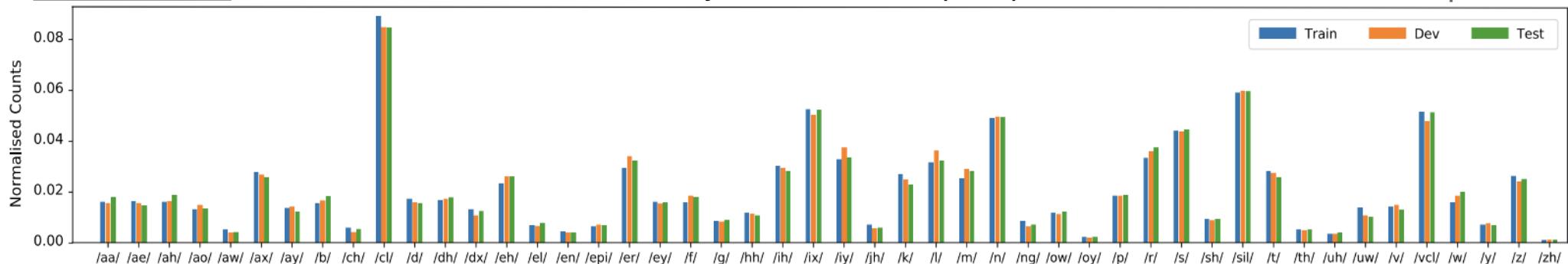


# Phonetic Distribution

Data: TIMIT

Probability Mass Function (PMF)

48 phones



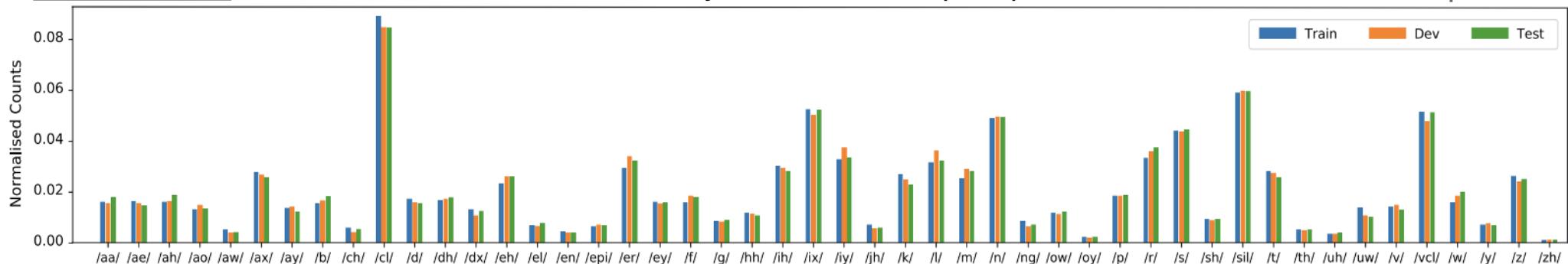
PMF of standard Train/Dev/Test sets is **identical**.

# Phonetic Distribution

Data: TIMIT

Probability Mass Function (PMF)

48 phones



PMF of standard Train/Dev/Test sets is **identical**.

PMF is **not uniform** ...

# Phonetic Distribution

Data: TIMIT

Probability Mass Function (PMF)

48 phones



PMF of standard Train/Dev/Test sets is **identical**.

PMF is **not uniform** ...

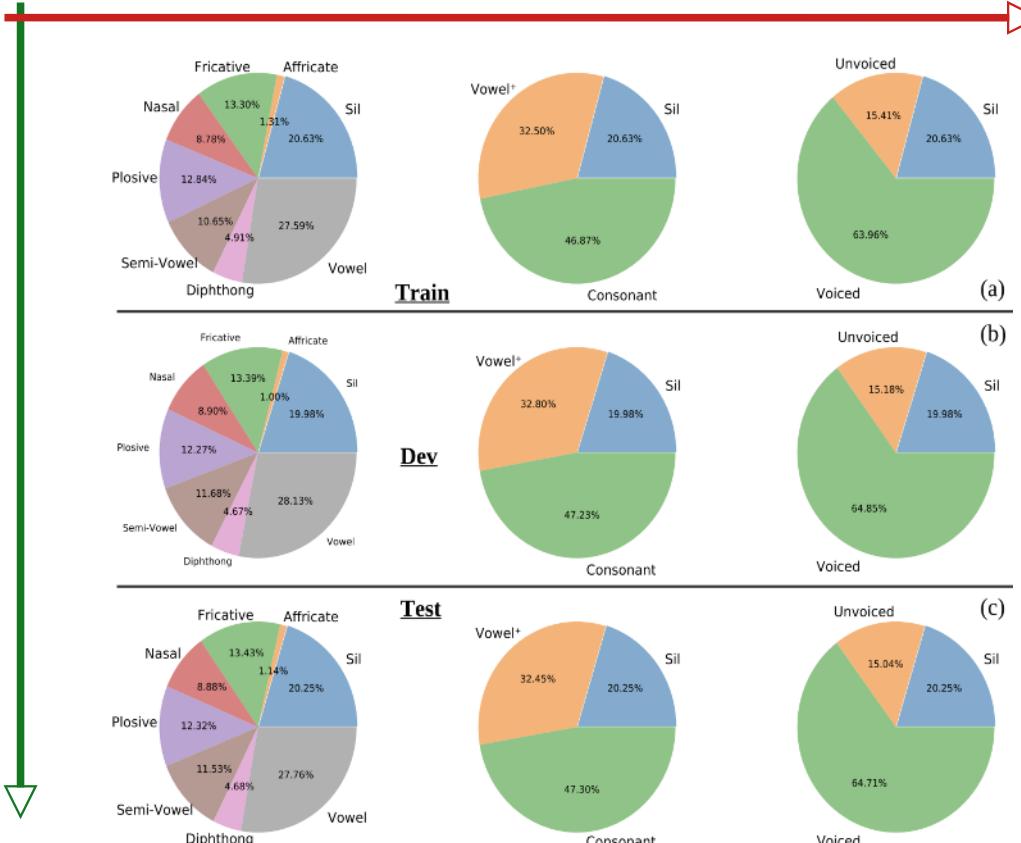
- ✗ Not perfect from *learning* perspective!
- ✓ Not a shortcoming, though → characteristic of natural languages, studied in *Quantal Theory, Adaptive Dispersion*, etc.

# PMF over BPCs

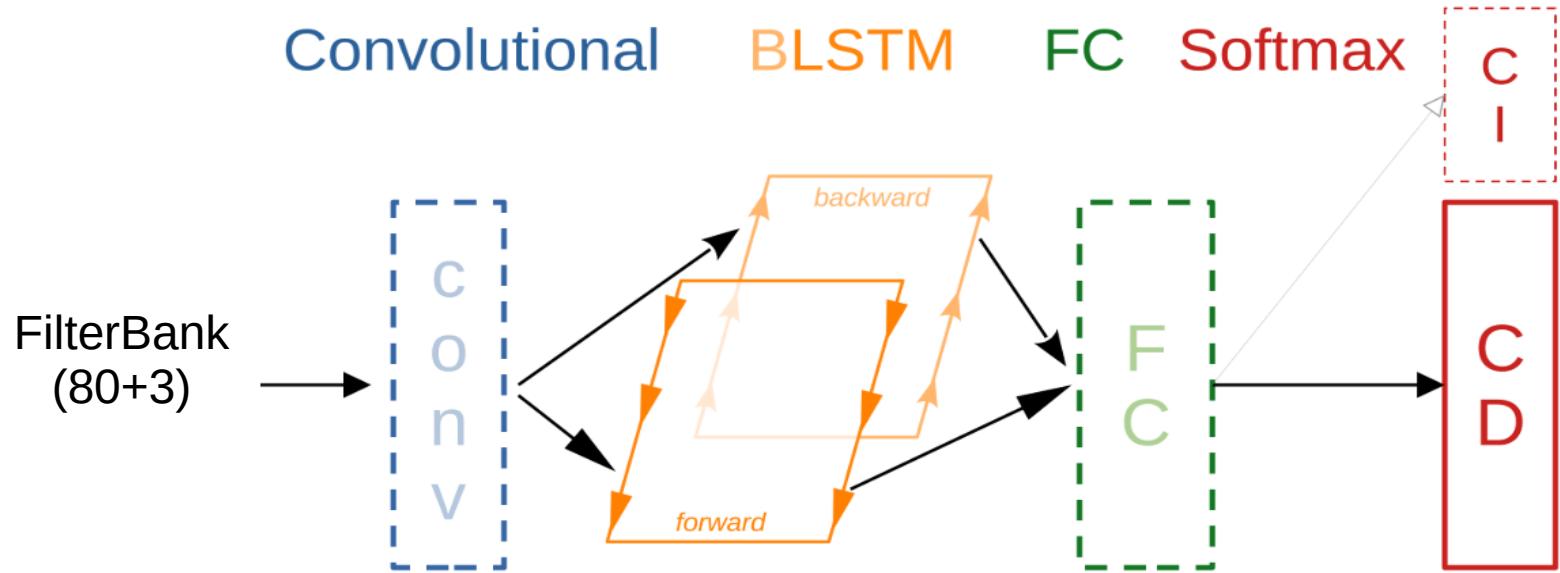
Non-Uniform inside BPC, e.g., Nasal 9%, Vowel 28%, Sil 21%, ...

Identical over  
Train/Dev/Test

e.g., Nasal: 8.8/8.9/8.9%



# Baseline Architecture



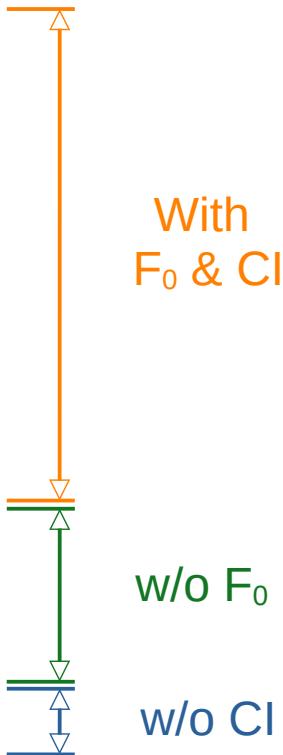
Architecture: **C<sub>i</sub>L<sub>j</sub>F<sub>k</sub>**

- $i$  convolutional layers
- $j$  BLSTM layers
- $k$  FC (fully-connected) layers

- Trained by cross entropy loss
- CI: Context Independent (48D)
- CD: Context Dependent (1936D)

# Choosing Baseline

Feature	Architecture	Dev	Test	#Param (M)
FBank-83	L2	13.1	15.2	7.2
FBank-83	L3	13.1	14.6	10.9
FBank-83	L4	<b>12.8</b>	<b>14.1</b>	14.5
FBank-83	L5	12.6	14.3	18.2
FBank-83	L6	13.0	15.0	21.8
FBank-83	L4F1	12.9	14.9	15.5
FBank-83	C1L4	12.7	14.4	20.9
FBank-83	C1L4F1	13.0	14.6	21.8
FBank-80	L4	12.8	14.3	14.5
FBank-40	L4	12.7	14.5	14.4
FBank-23	L4	13.2	14.5	14.3
FBank-83*	L4	13.0	14.6	14.4



# Choosing Baseline

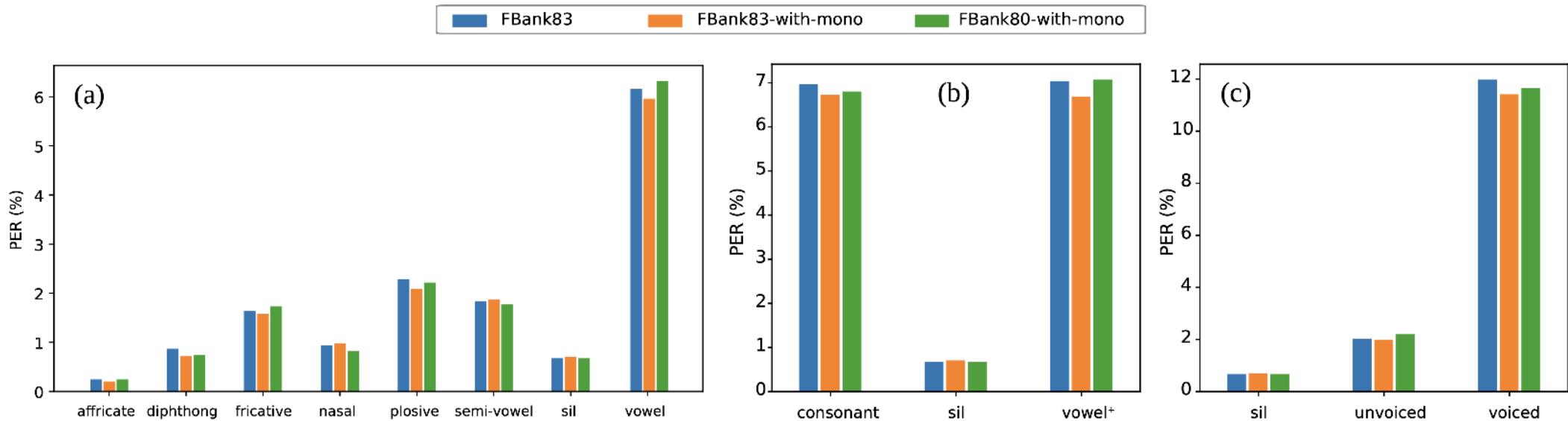
Feature	Architecture	Dev	Test	#Param (M)		
FBank-83	L2	13.1	15.2	7.2		
FBank-83	L3	13.1	14.6	10.9		
FBank-83	L4	<b>12.8</b>	<b>14.1</b>	14.5		
FBank-83	L5	12.6	14.3	18.2		
FBank-83	L6	13.0	15.0	21.8		
FBank-83	L4F1	12.9	14.9	15.5		
FBank-83	C1L4	12.7	14.4	20.9		
FBank-83	C1L4F1	13.0	14.6	21.8		
w/o $F_0$	→	FBank-80	L4	12.8	14.3	14.5
		FBank-40	L4	12.7	14.5	14.4
		FBank-23	L4	13.2	14.5	14.3
w/o CI	→	FBank-83*	L4	13.0	14.6	14.4

Chosen Baseline → With  $F_0$  & CI

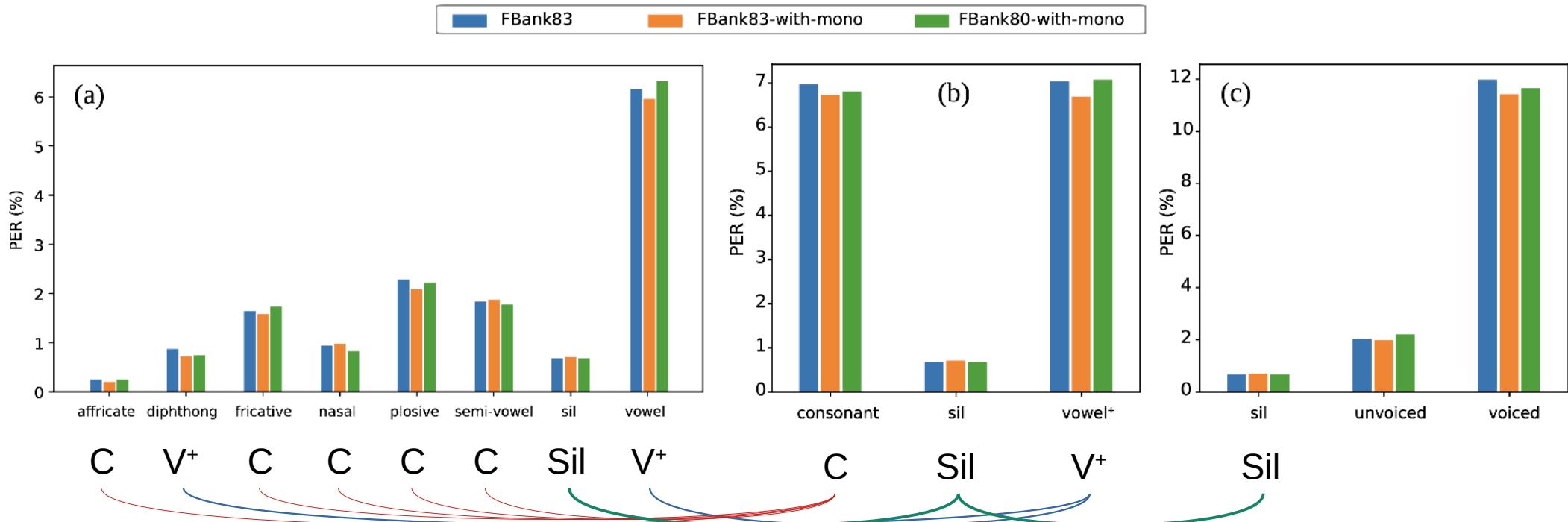
w/o  $F_0$

w/o CI

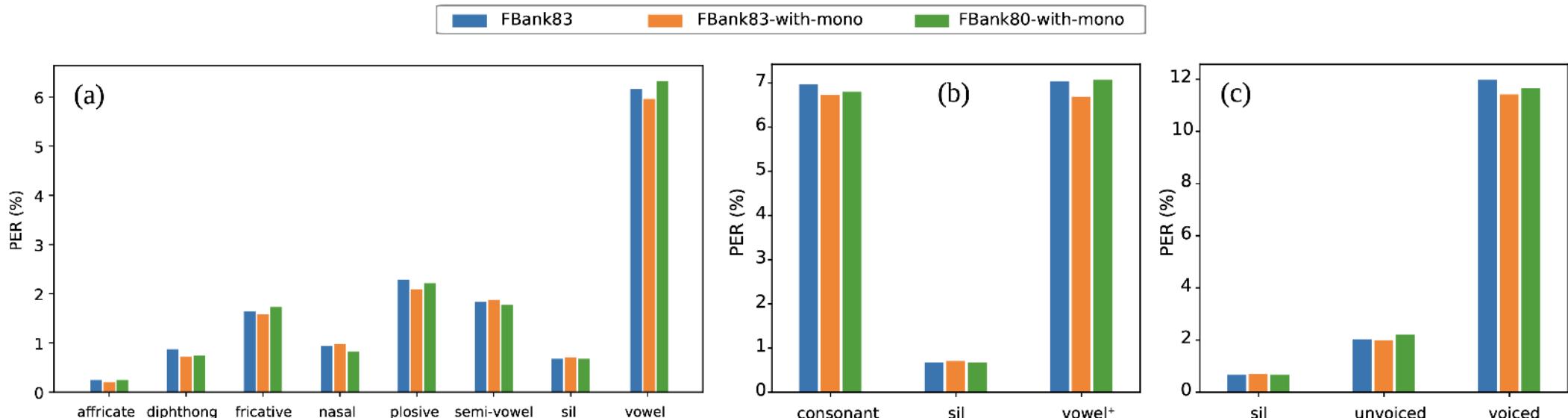
# Analysis beyond PER



# Analysis beyond PER

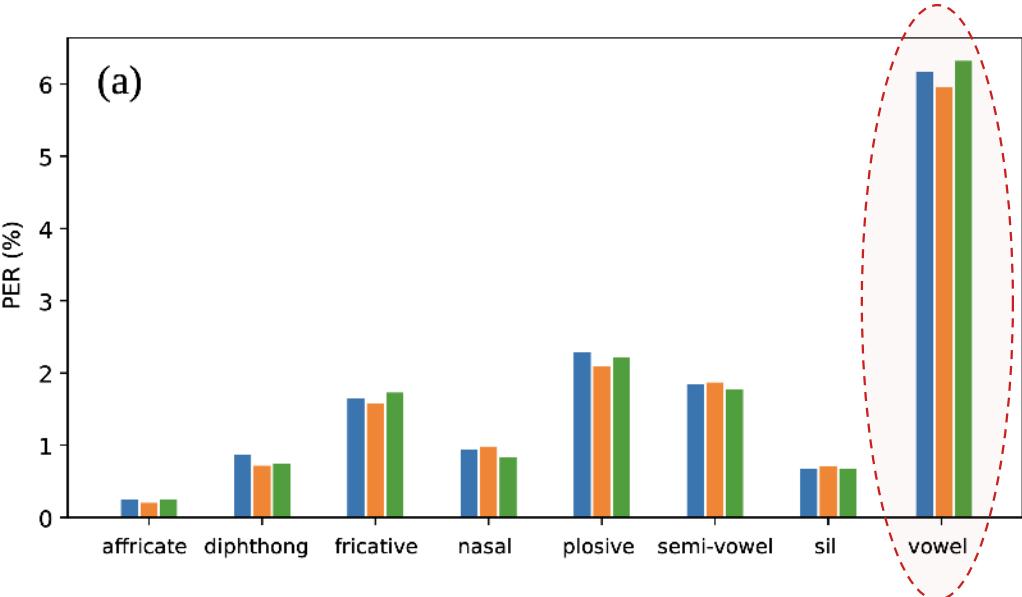


# Analysis beyond PER



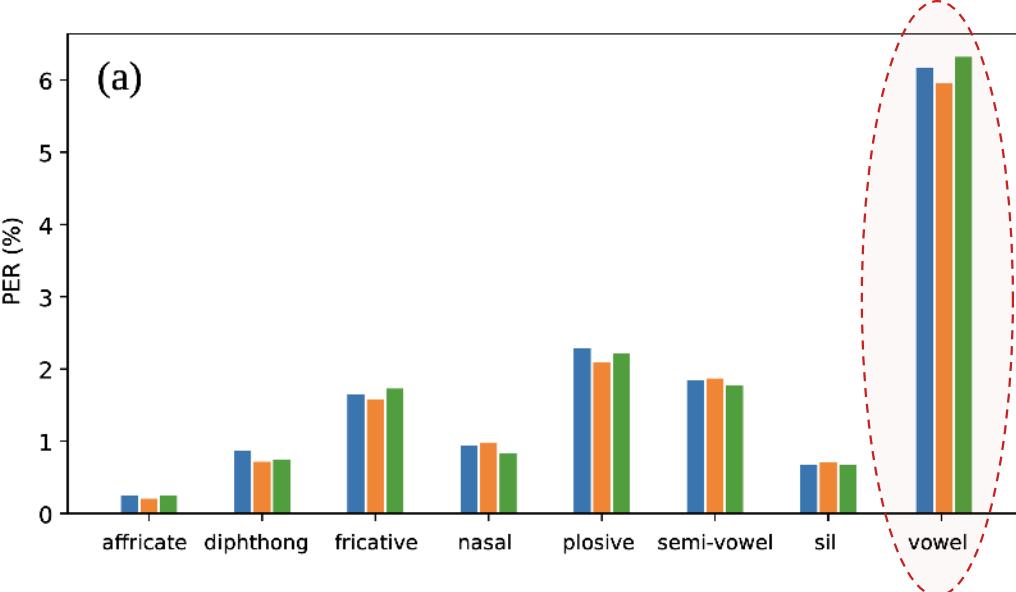
Minor gains after adding  $F_0$  features & regularisation with CI.

# Largest PER → Vowels



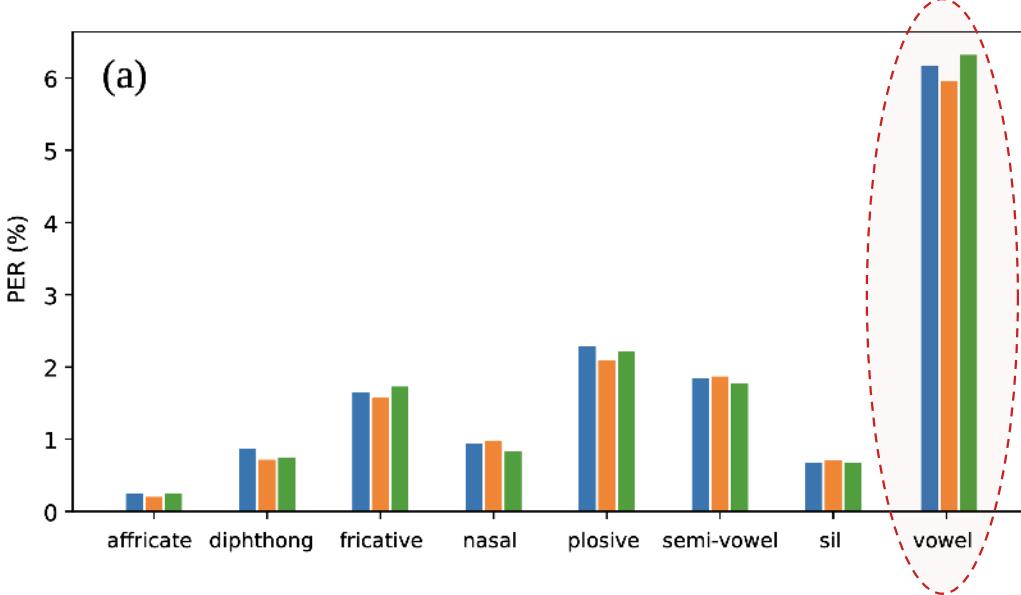
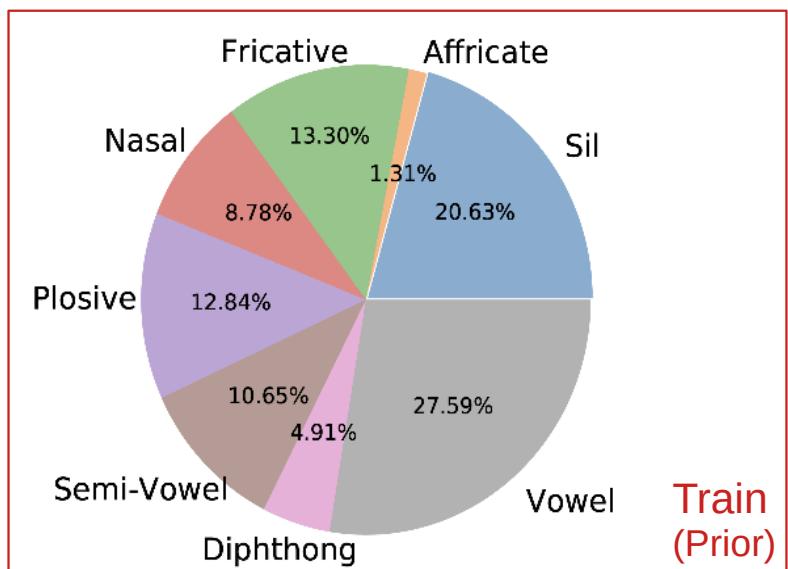
# Largest PER → Vowels

- Questions:
  - Training data amount?
  - TIMIT-specific?
  - Why?



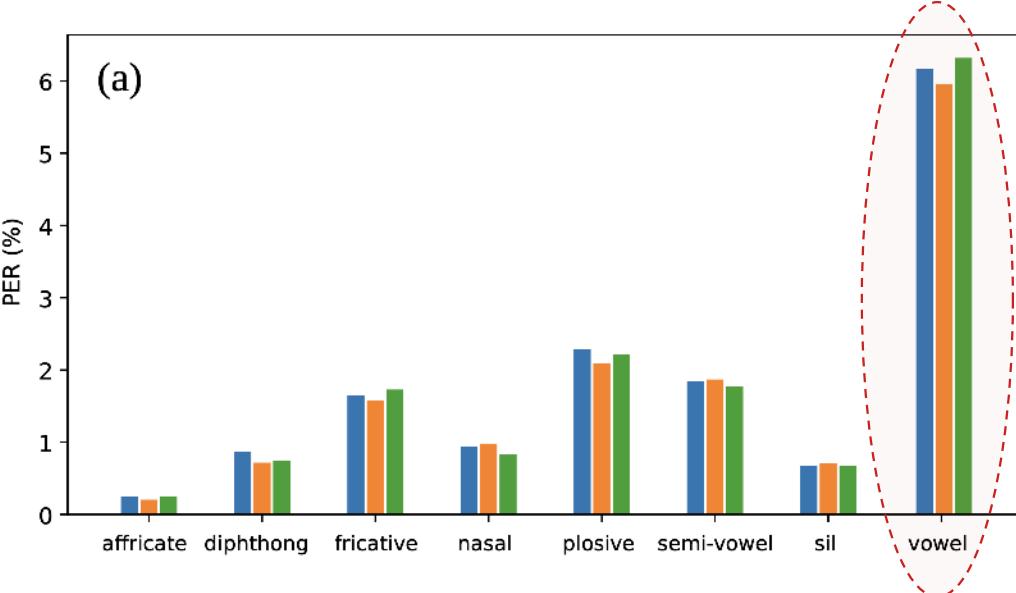
# Largest PER → Vowels

- Q1: Training data amount?



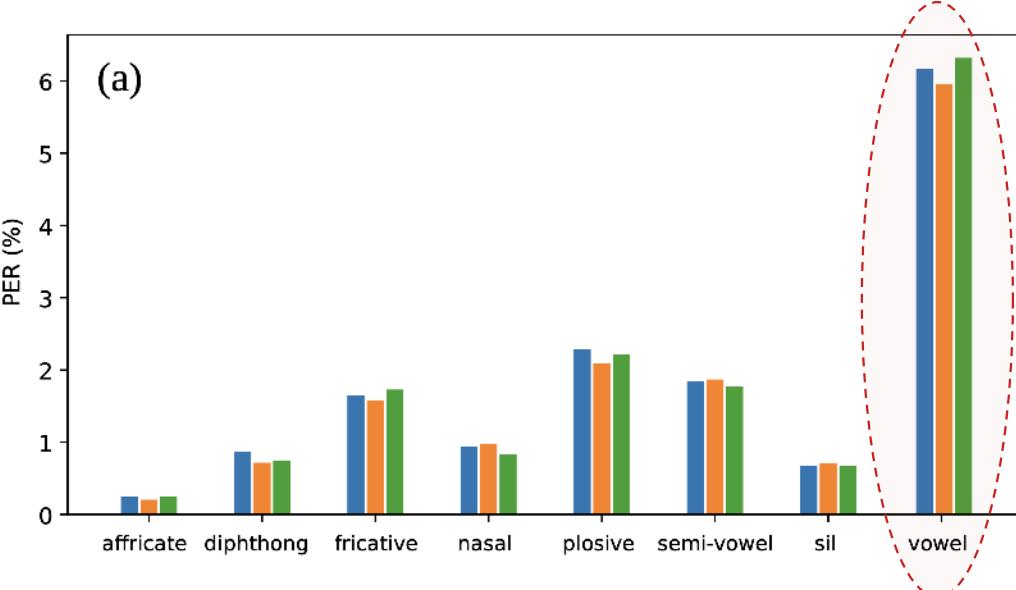
# Largest PER → Vowels

- Q2: TIMIT-specific?
  - Similar observation in human phone recognition [69]



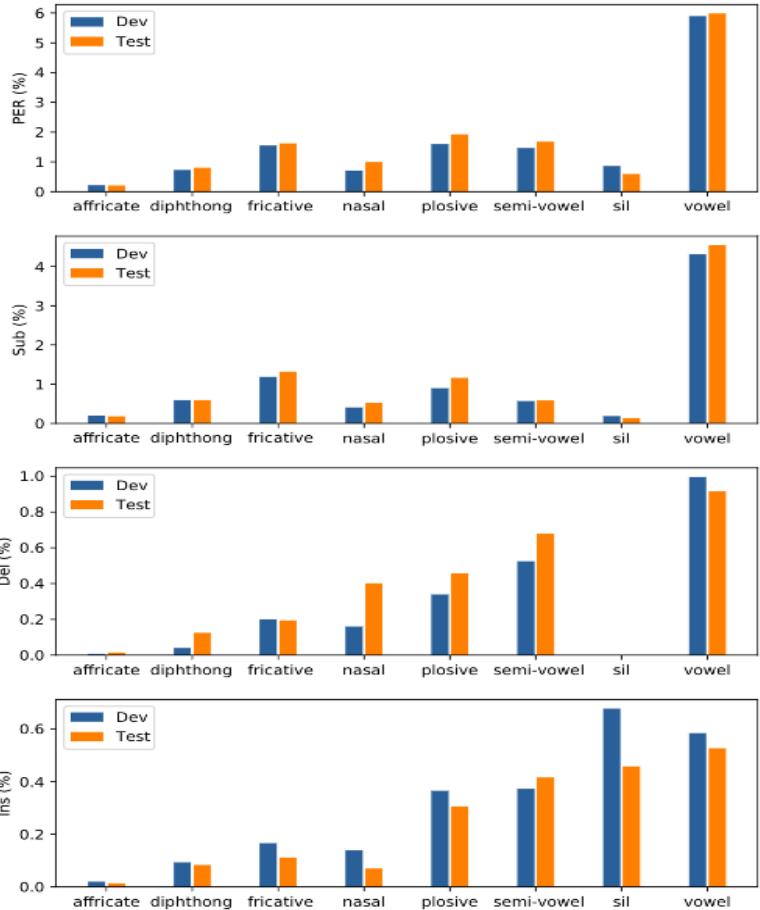
# Largest PER → Vowels

- Q2: TIMIT-specific?
  - Similar observation in human phone recognition [69]
- Q3: Why?

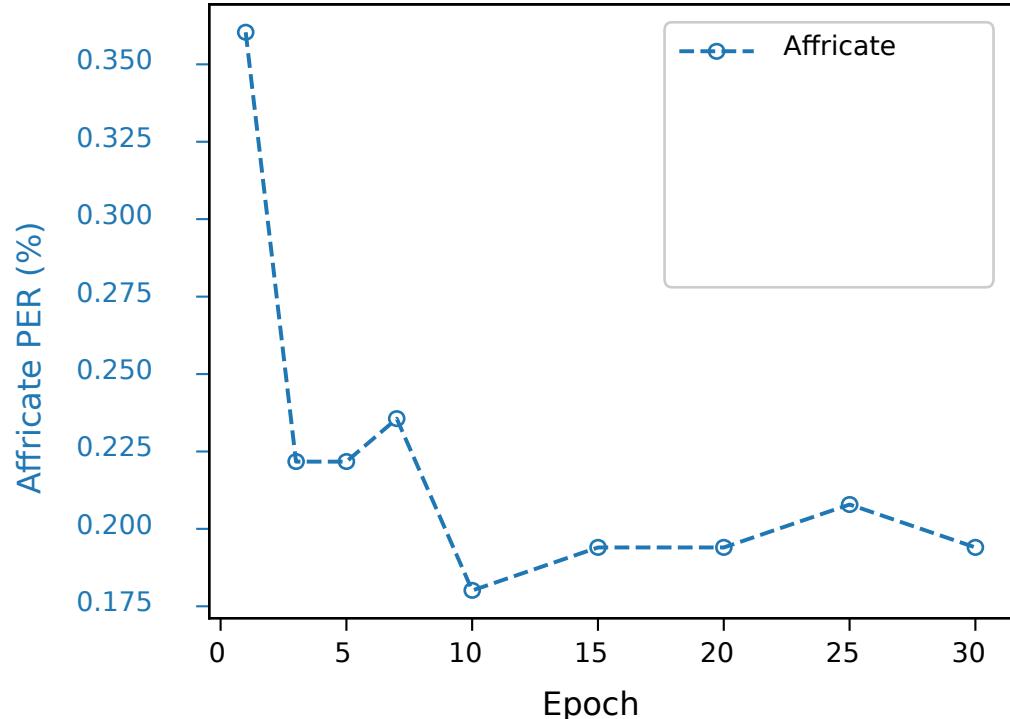


# Sub/Del/Ins per BPC

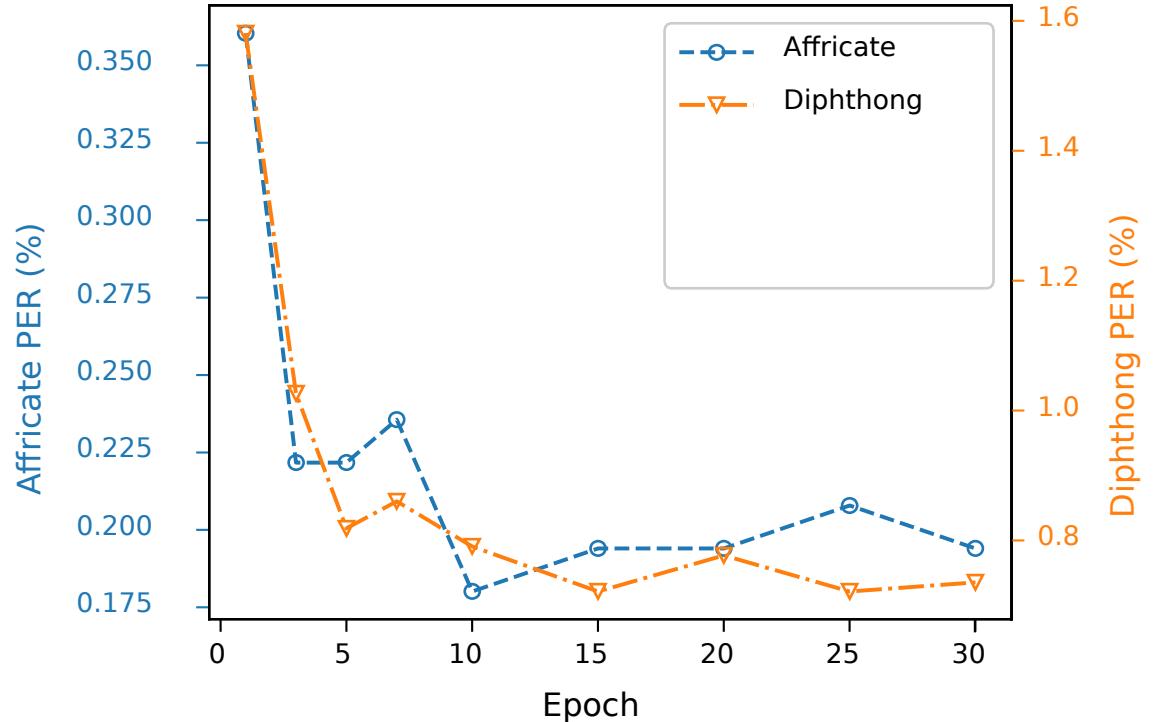
- Vowels
  - largest Sub/Del/Ins
- Silence
  - Small(est) Del
  - large(st) Ins



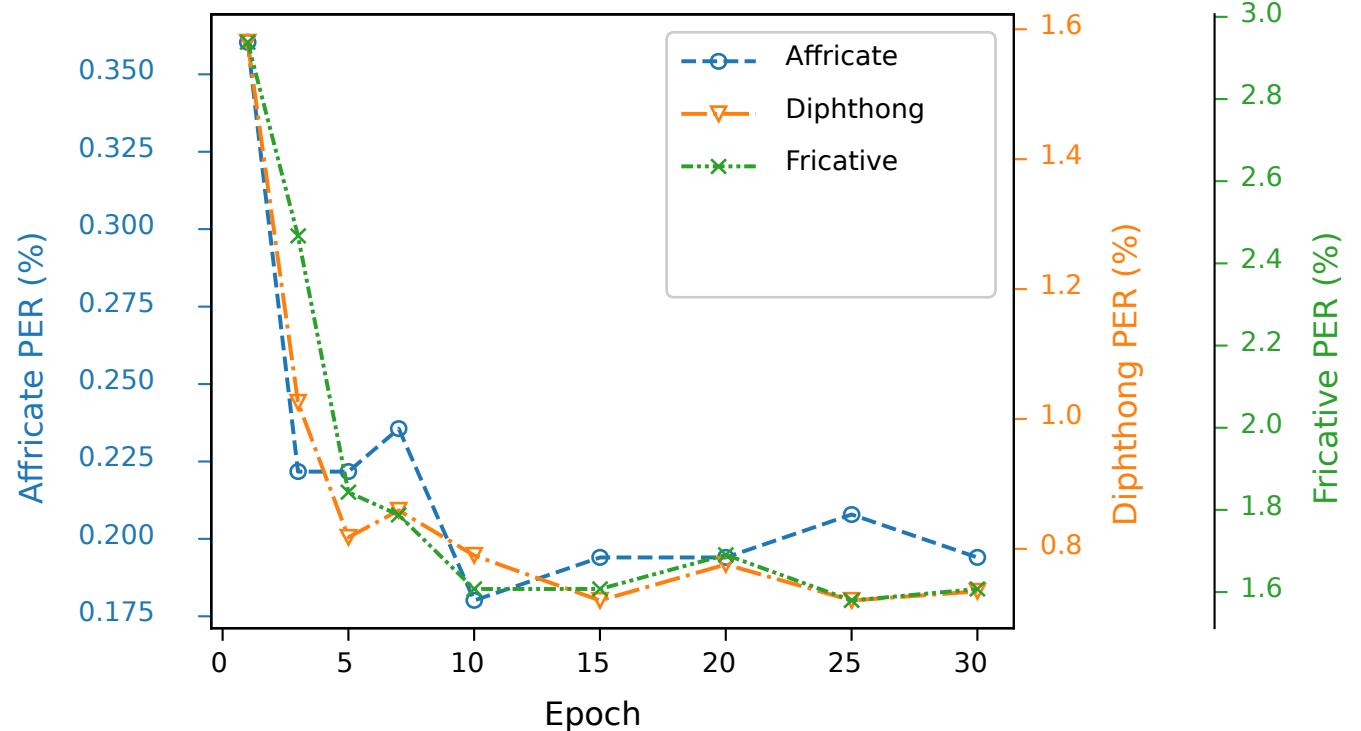
# Training Dynamics (1)



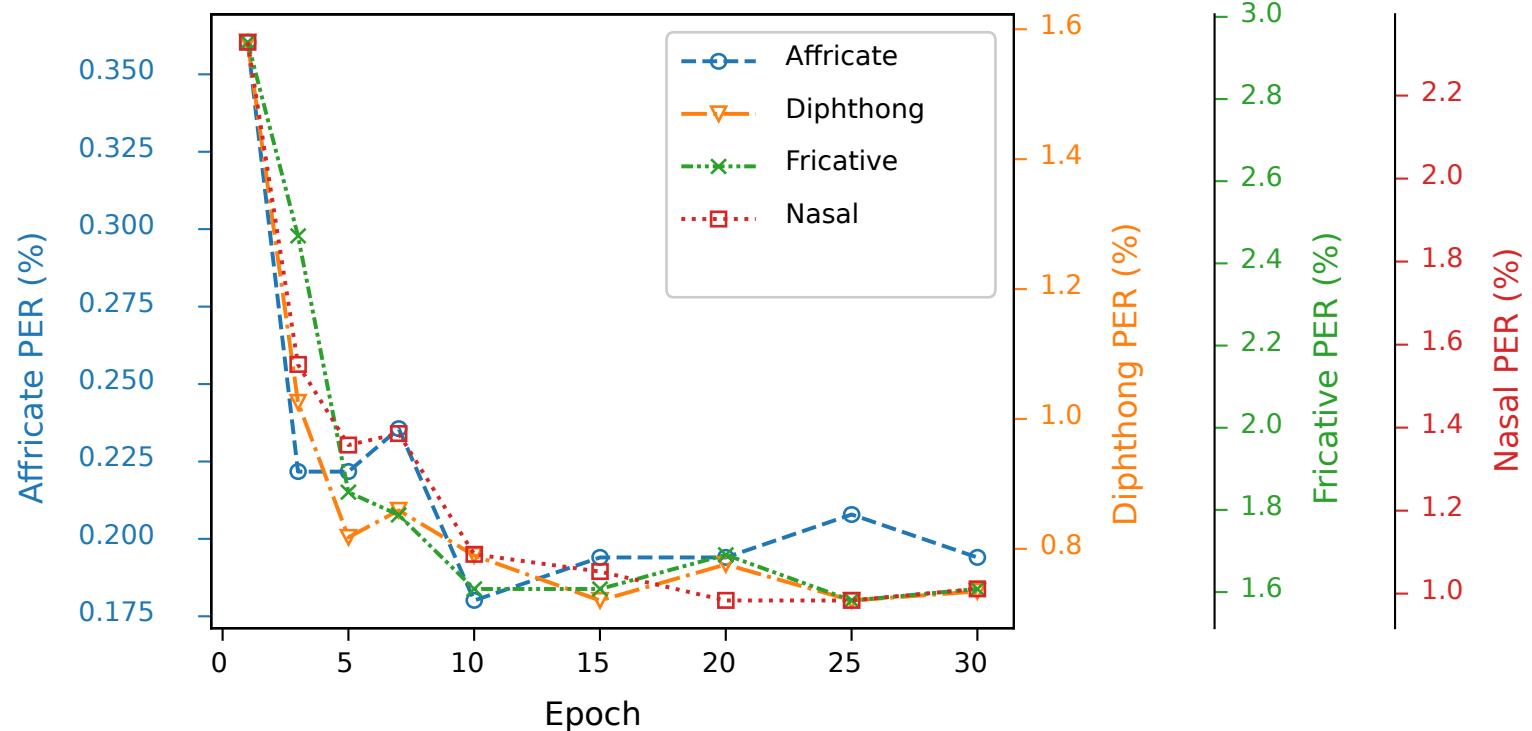
# Training Dynamics (1)



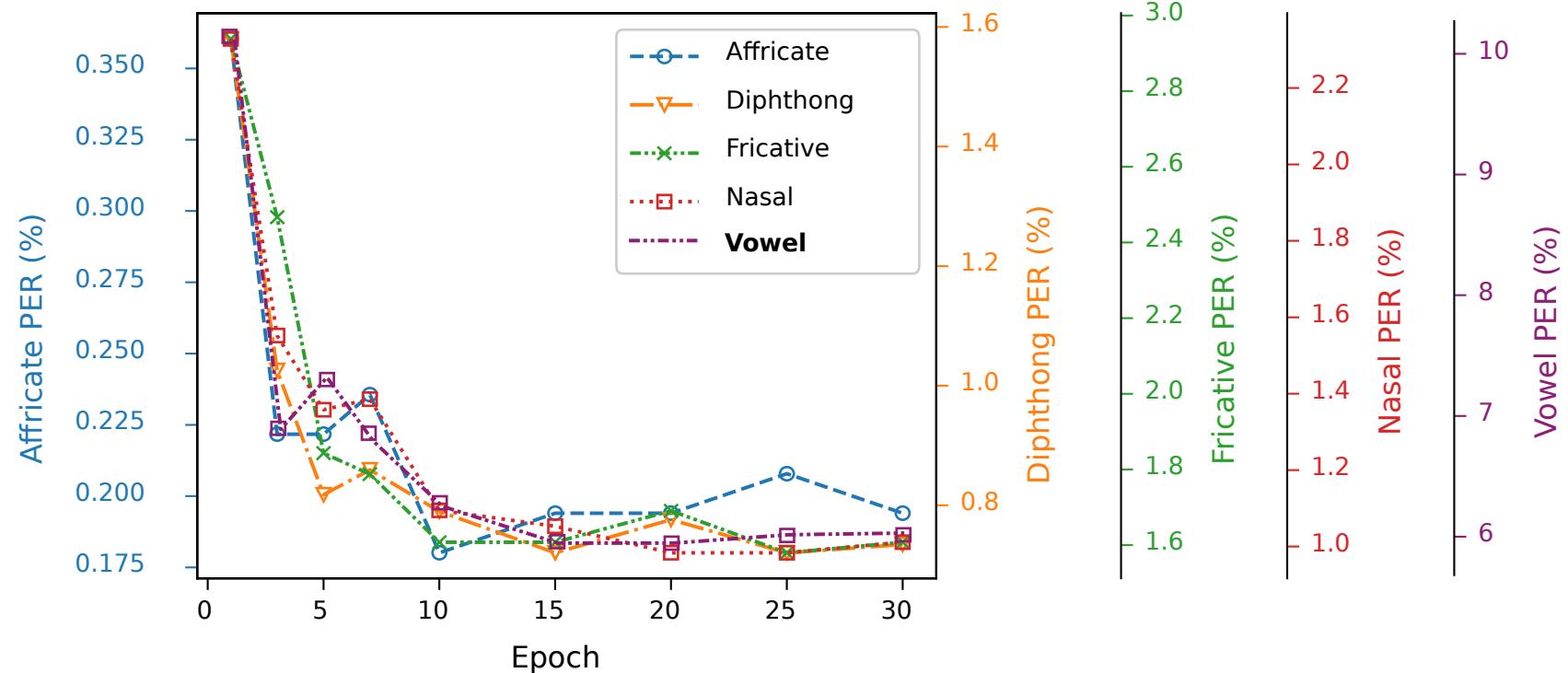
# Training Dynamics (1)



# Training Dynamics (1)



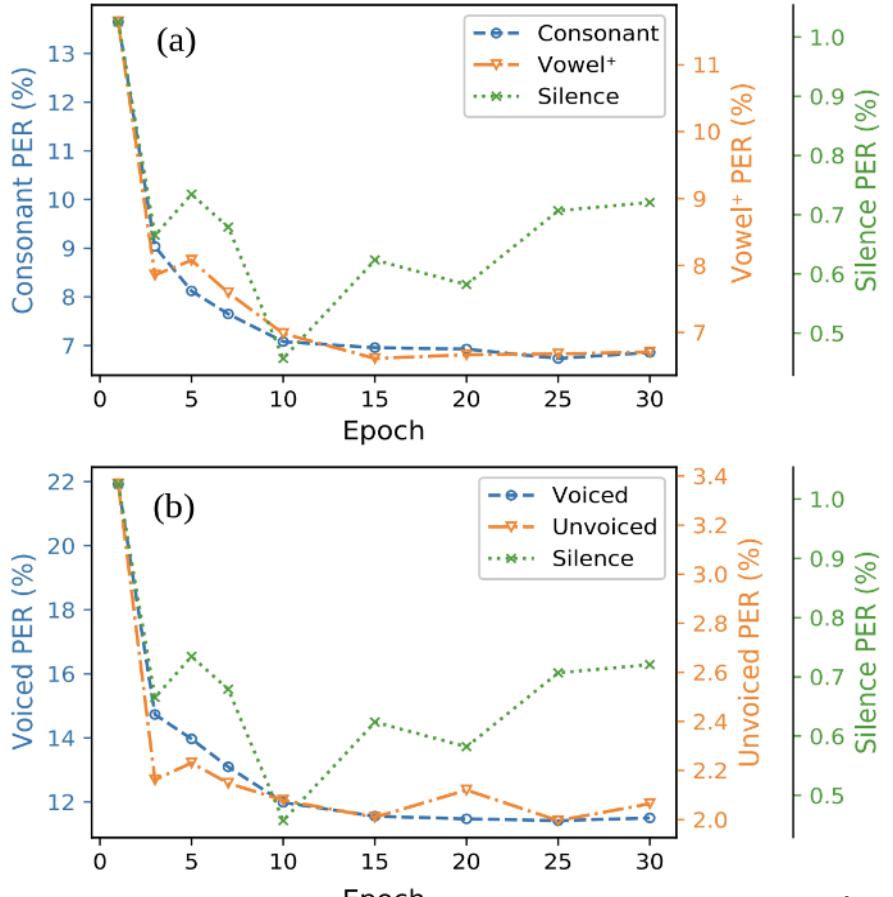
# Training Dynamics (1)



# Training Dynamics (2)



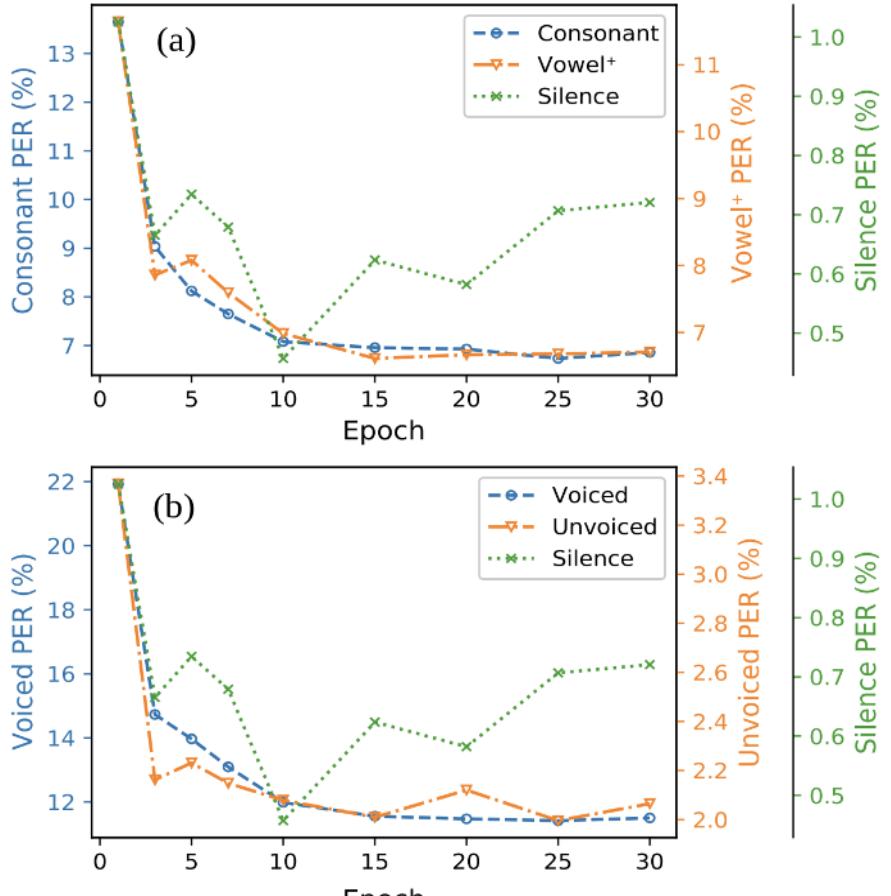
# Training Dynamics (3)



# Training Dynamics (3)

Similar dynamics for all classes;  
despite different PER (except Silence).

Dynamics is not class-specific;  
depends on architecture, loss and data.



# Confusion Matrices

Confusion matrices are computed using **Sub** errors.

The **bold** & underlined indicate the 1<sup>st</sup> & 2<sup>nd</sup> mostly confused classes.

True Label	aff	dip	fri	nas	plo	sem	sil	vow
aff	<b>10</b>	0	6	0	4	0	0	0
dip	0	<u>13</u>	0	1	1	<u>13</u>	0	<b>50</b>
fri	8	3	<b>127</b>	1	<u>24</u>	9	7	4
nas	0	1	3	<b>41</b>	<u>9</u>	4	3	5
plo	8	0	<u>25</u>	2	<b>73</b>	4	0	5
sem	5	16	12	3	7	<u>18</u>	2	<b>54</b>
sil	0	0	<b>4</b>	<u>4</u>	3	2	0	1
vow	1	<u>48</u>	4	5	5	<b>48</b>	3	<b>549</b>

(a)	Predicted Label								
(b)	sil con vow <sup>+</sup>			sil unv voi					
sil	0	<b>13</b>	<u>1</u>	sil	0	2	<b>12</b>		
con	12	<b>403</b>	<u>88</u>	unv	5	<u>55</u>	<b>84</b>		
vow <sup>+</sup>	3	<u>78</u>	<b>660</b>	voi	10	<u>125</u>	<b>965</b>		

# Confusion Matrices

**Fricatives are MCW Fricatives & Plosives.**

**Semi-vowels are MCW Vowels & Semi-vowels.**

**Silence is MCW Fricatives & Nasals.**

MCW: mostly confused with

⋮

	aff	dip	fri	nas	plo	sem	sil	vow
aff	10	0	6	0	4	0	0	0
dip	0	13	0	1	1	13	0	50
fri	8	3	127	1	24	9	7	4
nas	0	1	3	41	9	4	3	5
plo	8	0	25	2	73	4	0	5
sem	5	16	12	3	7	18	2	54
sil	0	0	4	4	3	2	0	1
vow	1	48	4	5	5	48	3	549

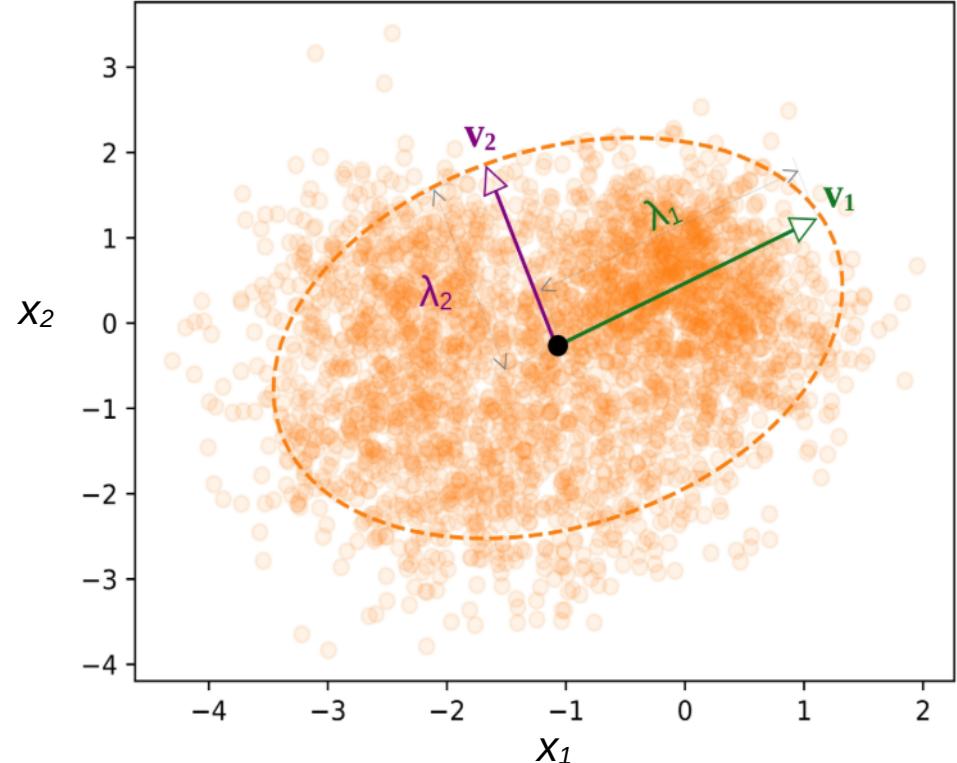
	Predicted Label		
(a)			
(b)	sil	con	vow <sup>+</sup>
sil	0	13	1
con	12	403	88
vow <sup>+</sup>	3	78	660

	sil	unv	voi
(c)	sil	unv	voi
sil	0	2	12
unv	5	55	84
voi	10	125	965

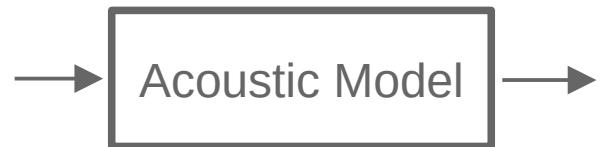
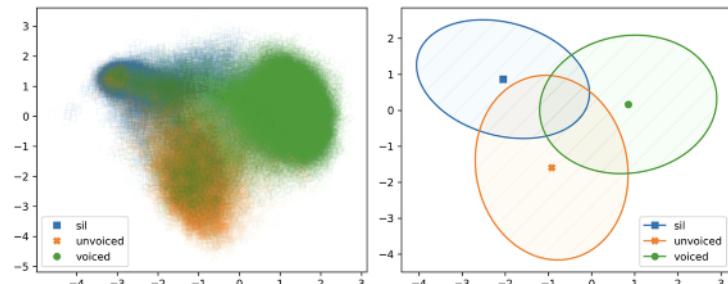
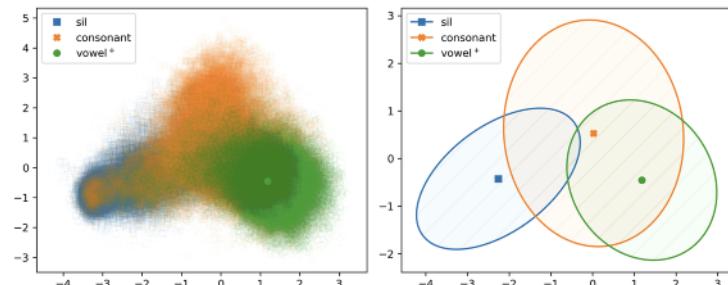
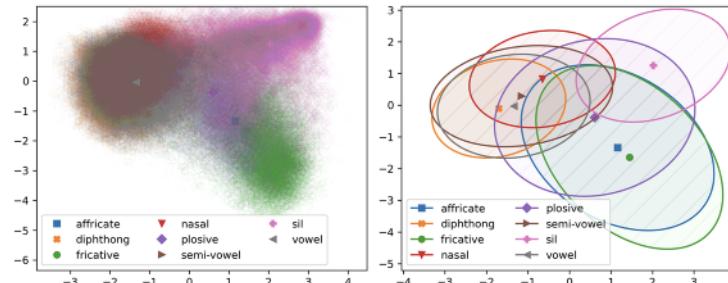
# Scatter Plot in 2D

- How:
  - LDA (t-SNE later)
  - Fit an ellipse (How?)

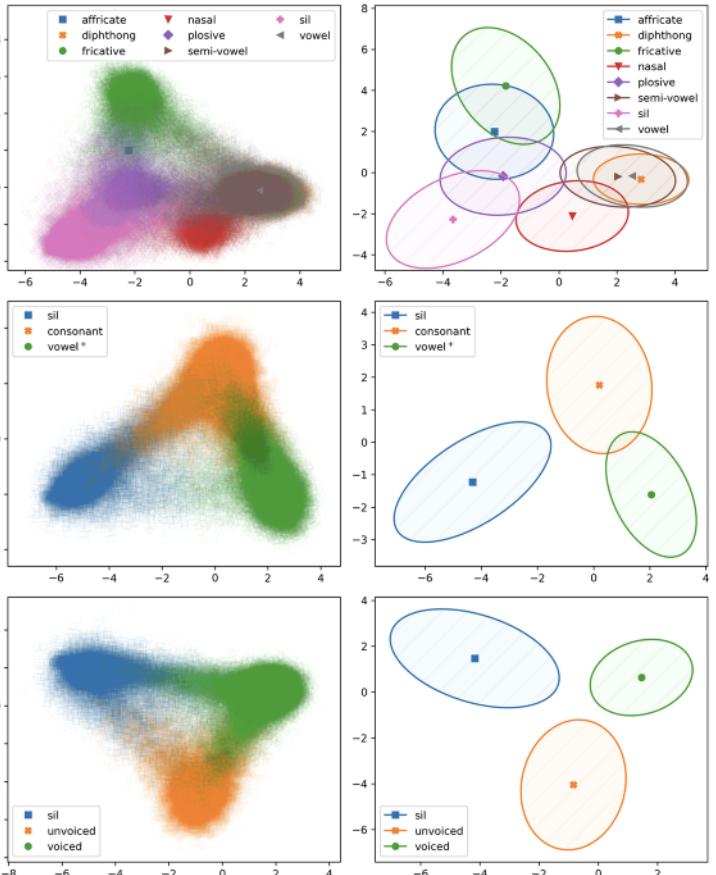
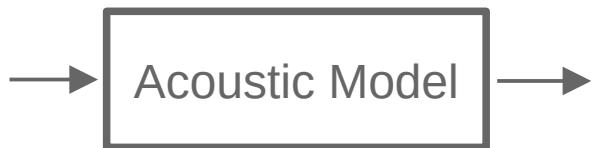
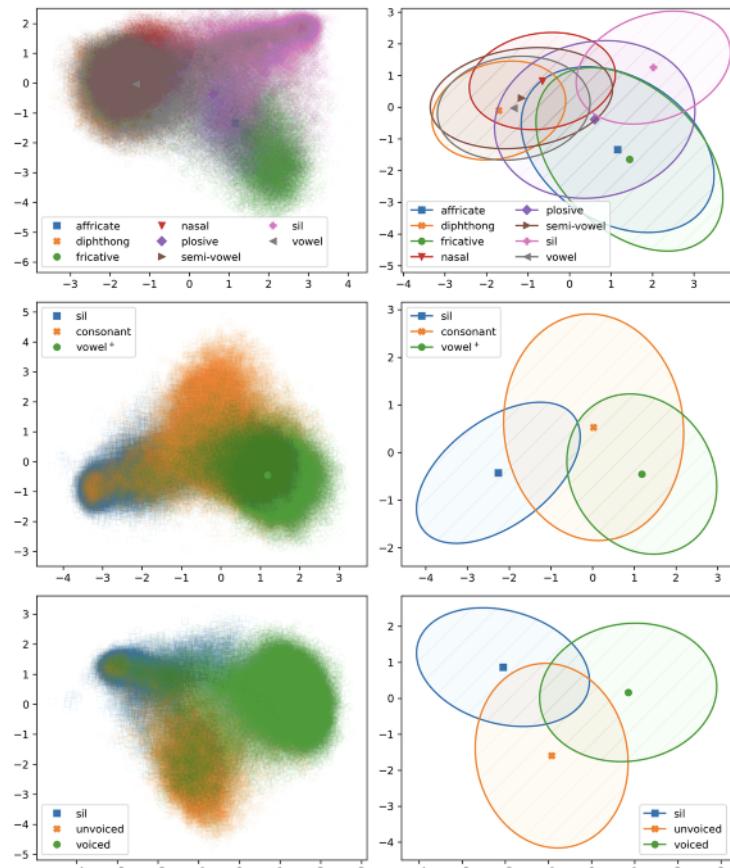


\* LDA: Linear Discriminant Analysis

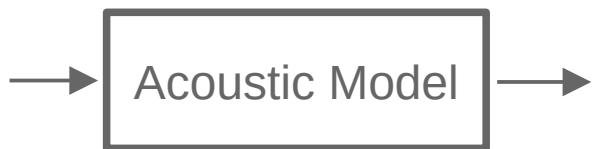
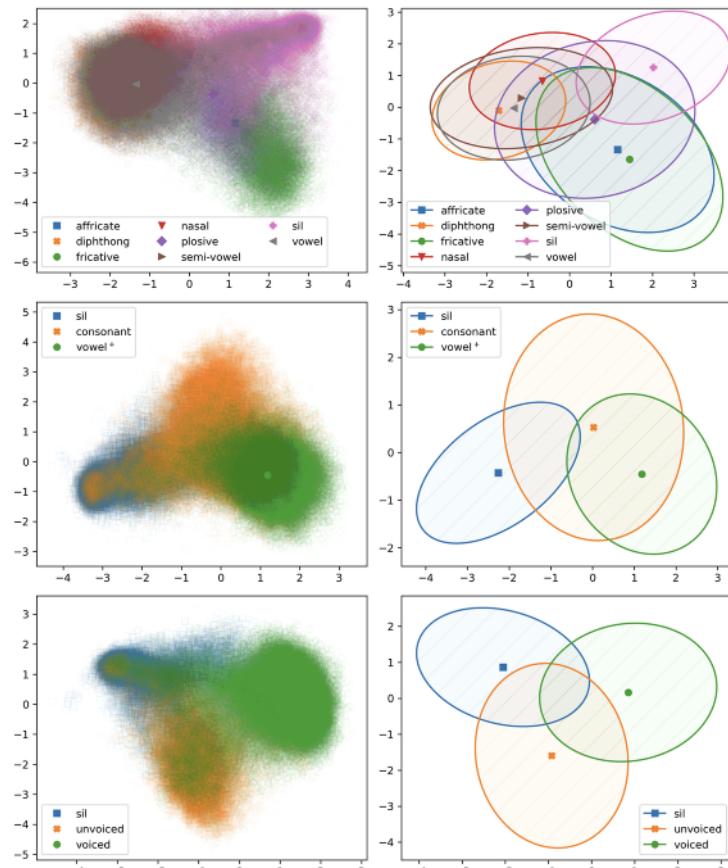
# Scatter Plots



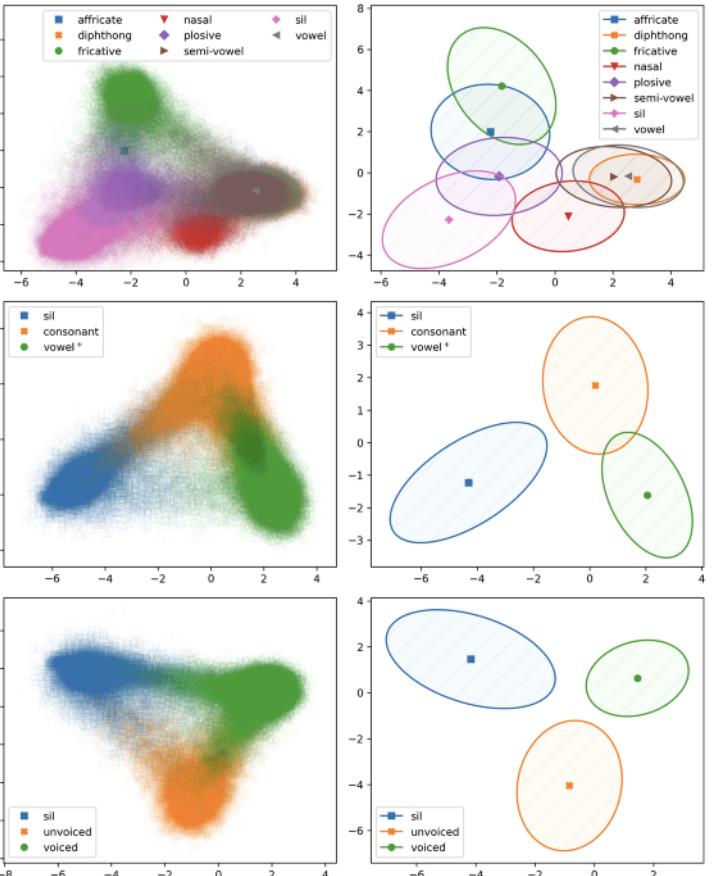
# Scatter Plots



# Scatter Plots



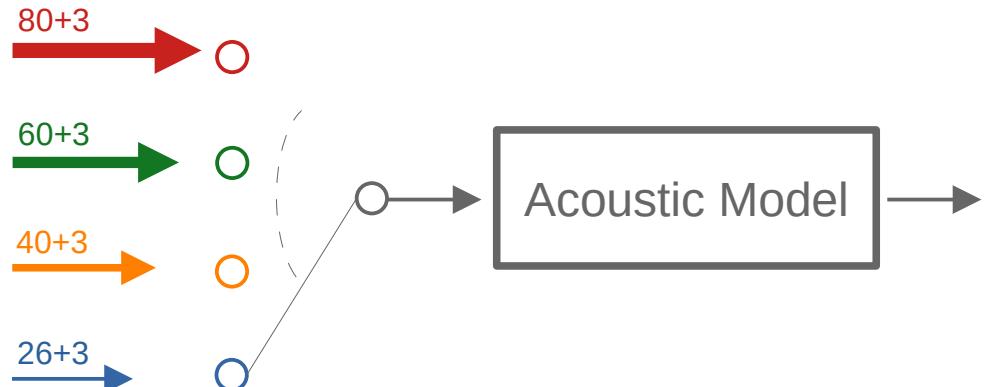
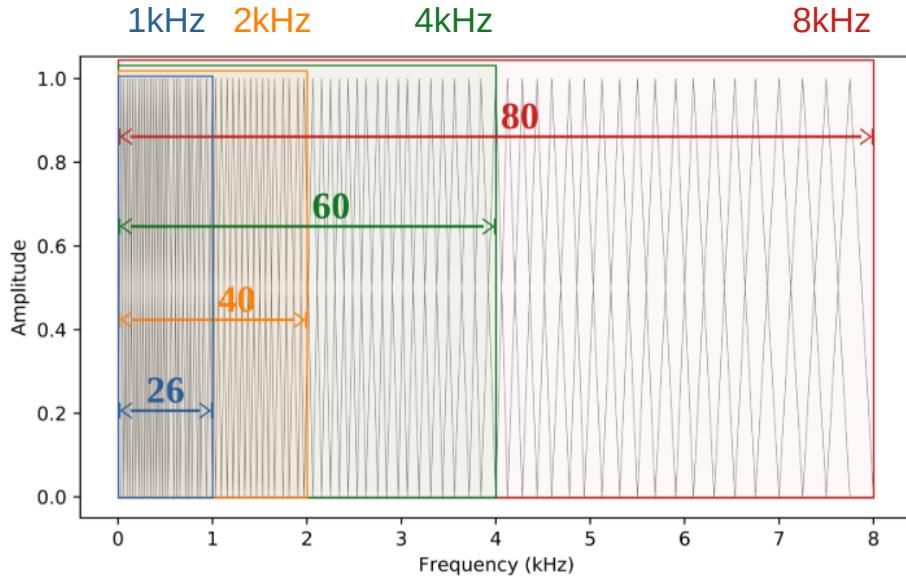
- ✓ → More distinct clusters →
- ✗ Scatter plots justify confusions (but not perfectly!)



# Various Systems

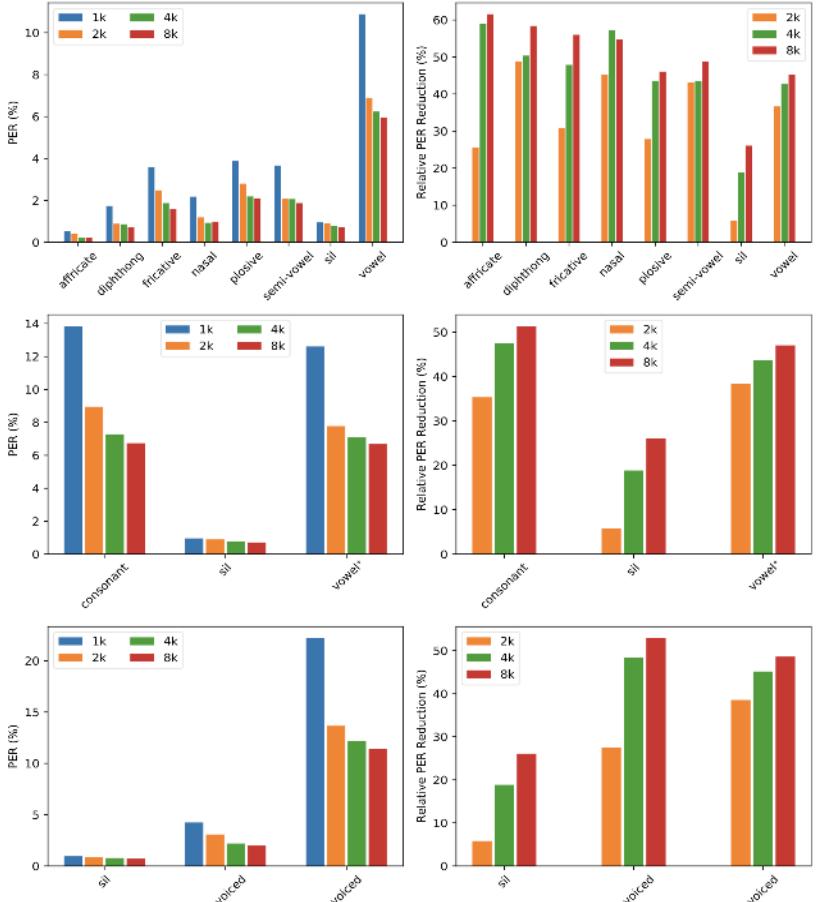
Model	Task	Architecture	Dev	Test
Baseline	TIMIT	L4-Hybrid	12.8	14.1
Subband-1k	TIMIT	L4-Hybrid	25.1	27.3
Subband-2k	TIMIT	L4-Hybrid	16.8	17.6
Subband-4k	TIMIT	L4-Hybrid	13.4	15.0
UniLSTM	TIMIT	L4-Hybrid	15.9	17.8
Baseline	NTIMIT	L4-Hybrid	19.2	20.1
GMM-HMM	TIMIT	SAT-MLLT-LDA	20.5	21.5
Baseline (WSJ*)	TIMIT	L4-Hybrid	11.5	13.1
Conformer	TIMIT	E2E	18.2	20.0
wav2vec 2.0	TIMIT	E2E (pre-trained)	7.1	8.3

# Effect of Sub-bands



# Effect of Sub-bands

Relative gain computed w.r.t. 1 kHz system.

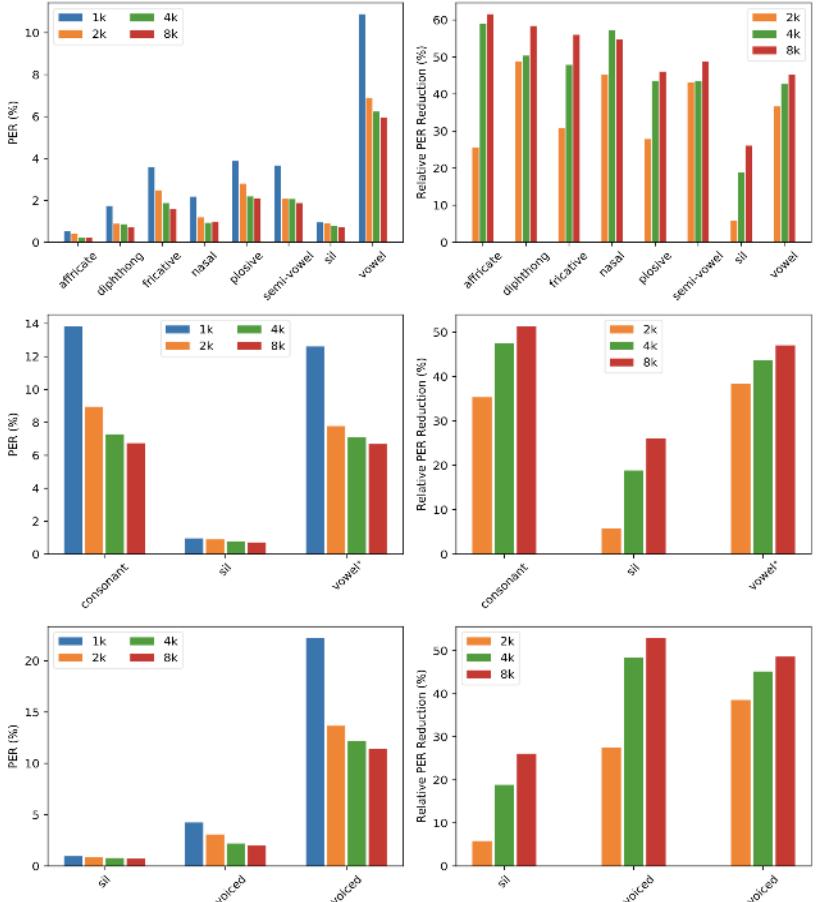


# Effect of Sub-bands

Relative gain computed w.r.t. 1 kHz system.

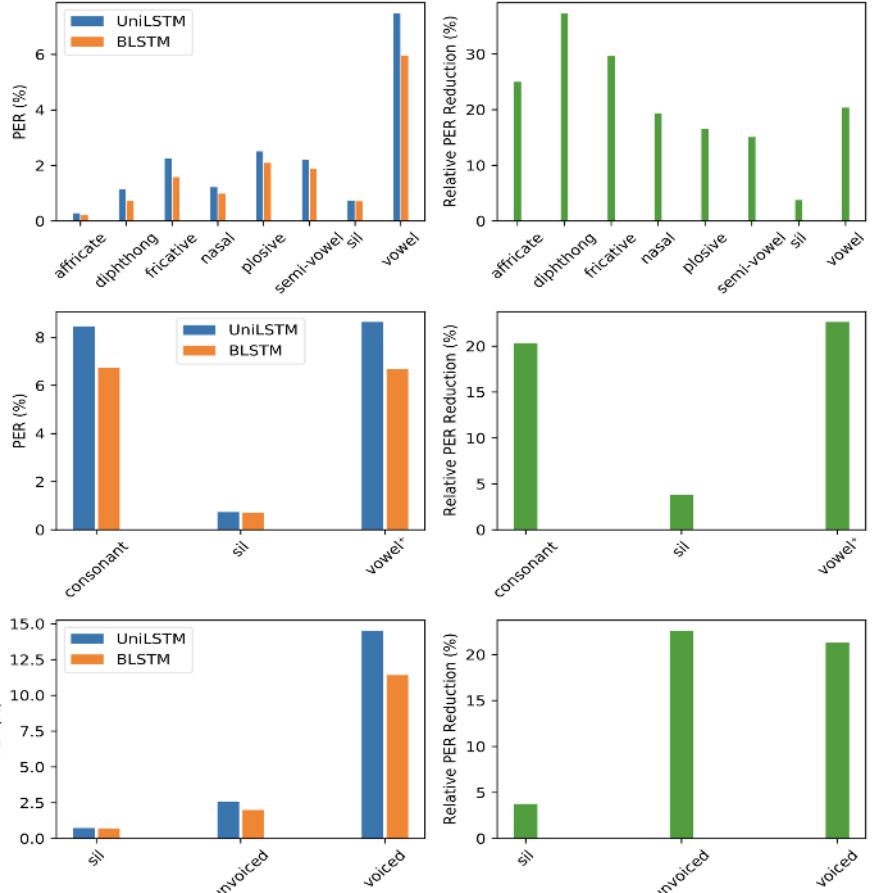
Including freq > 2kHz ...

- Small yet consistent gain for Voiced, Semi/V<sup>+</sup>.
- Notable gain for Unvoiced, Aff/Fri/Nas/Plo/Sil



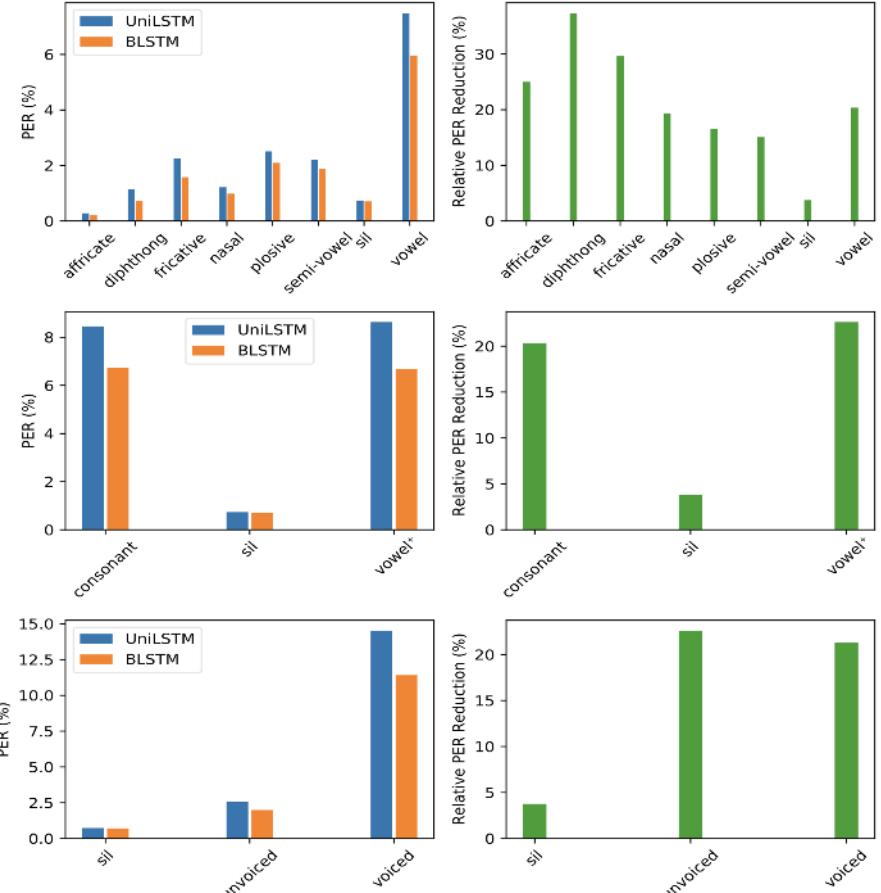
# Uni- vs Bi-Directional

- Relative Gain:
  - Typically 15% to 40%



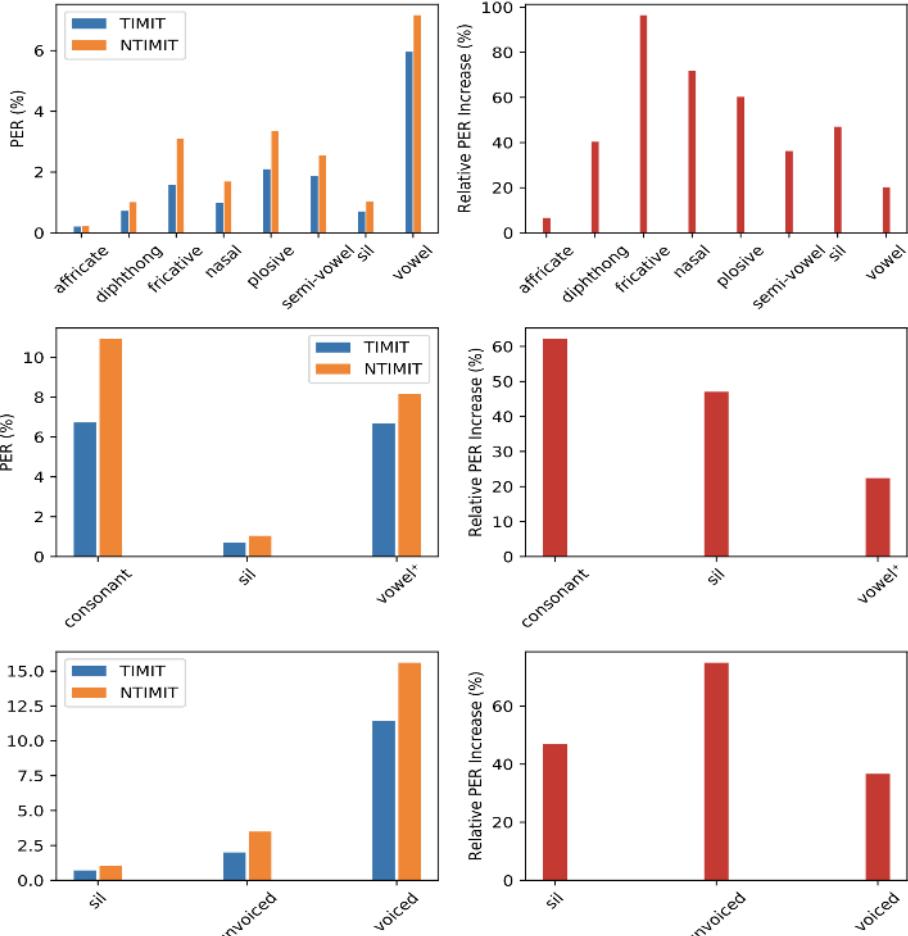
# Uni- vs Bi-Directional

- Relative Gain:
  - Typically 15% to 40%
- Silence benefits the least (4%)
  - Why?



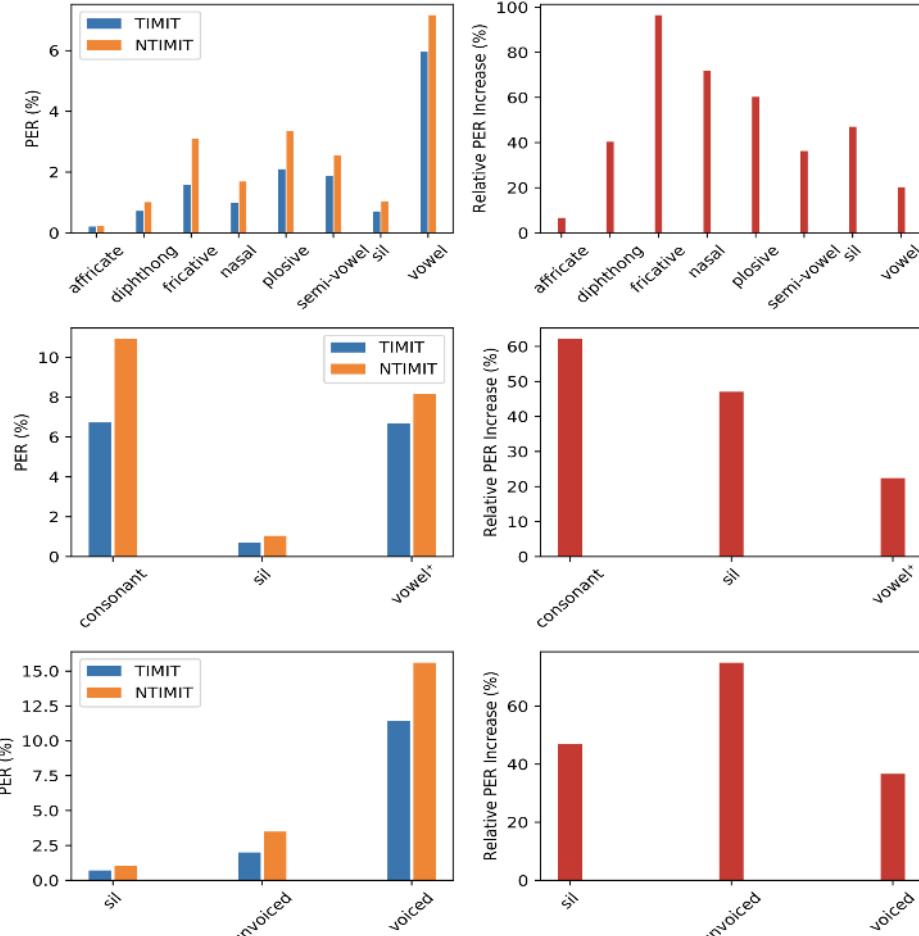
# NTIMIT

- Relative gain: -21% to -95%



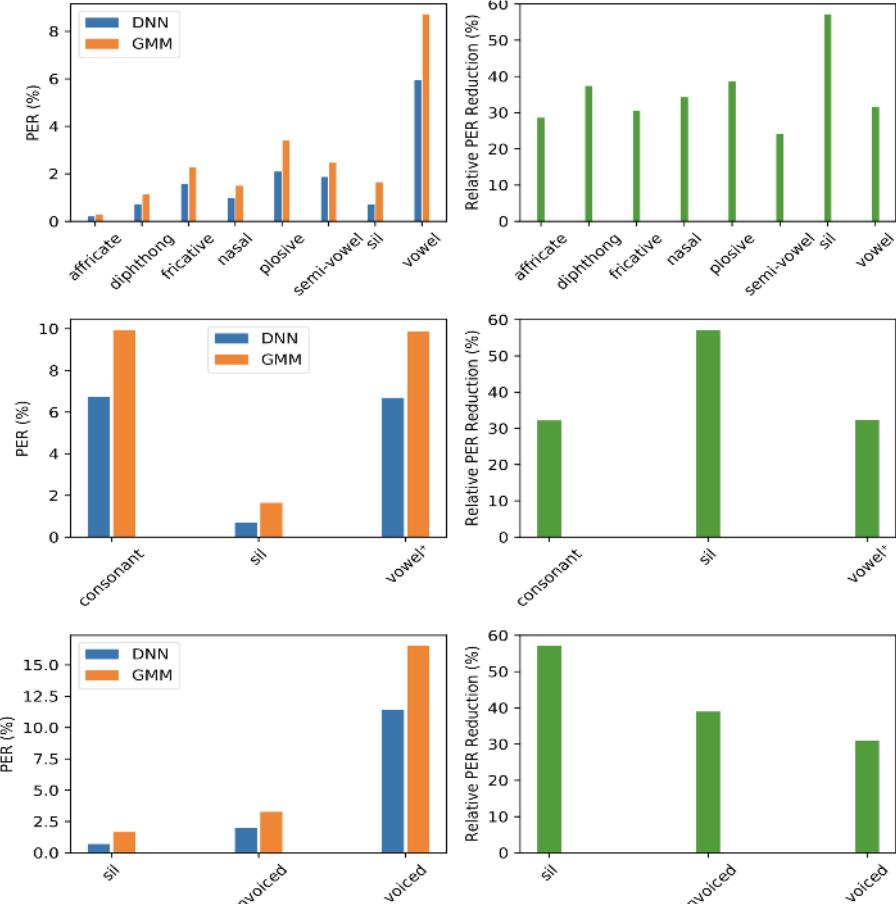
# NTIMIT

- Relative gain: -21% to -95%
- Robustness
  - Most: Vowels
  - Least: Fricatives
- Why?



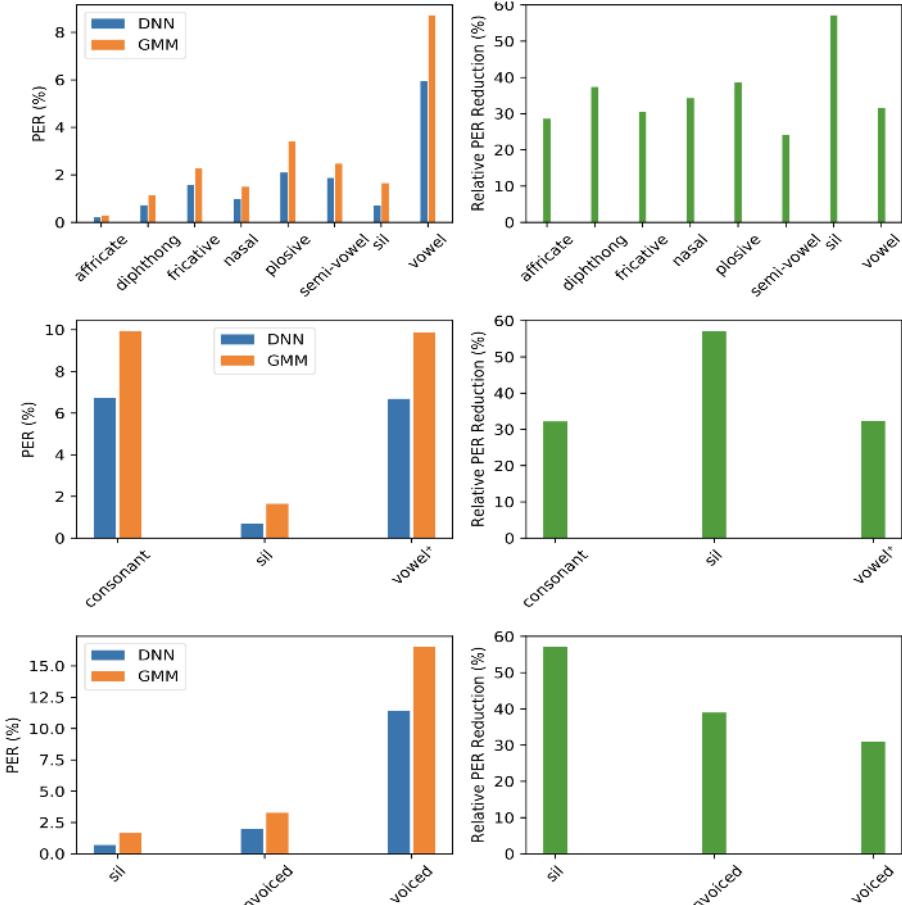
# GMM-HMM VS DNN-HMM

- Relative Gain: 25% to 55%



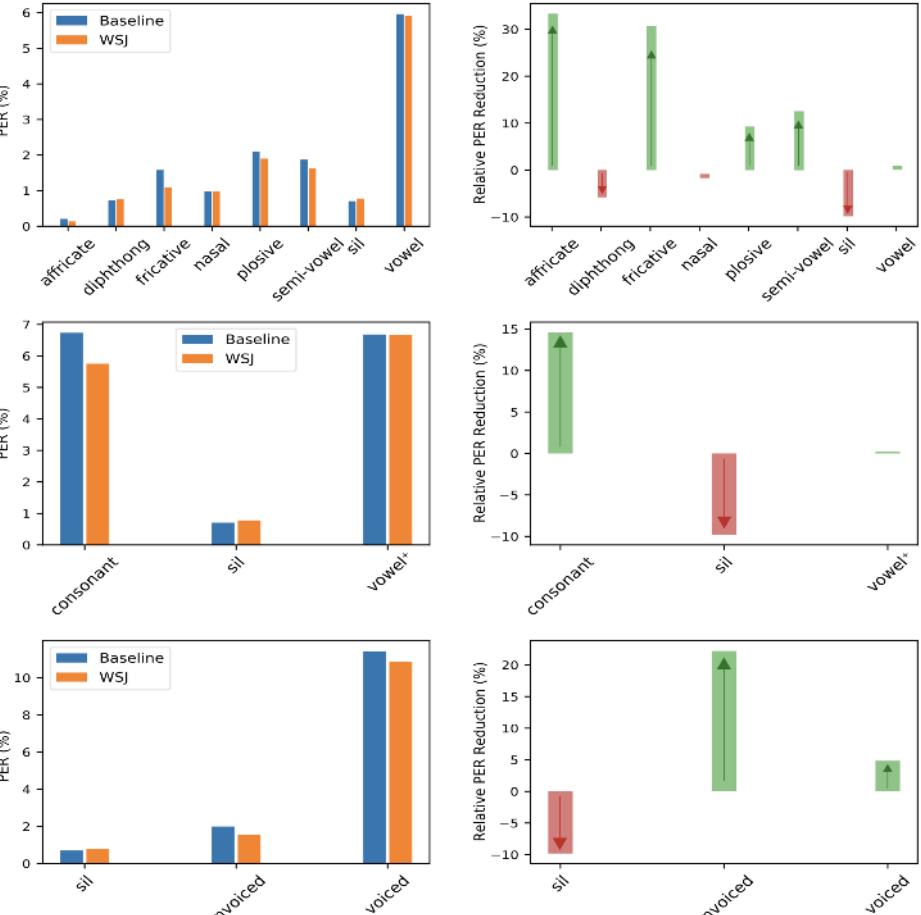
# GMM-HMM VS DNN-HMM

- Relative Gain: 25% to 55%
- Silence benefits the most
  - Why?



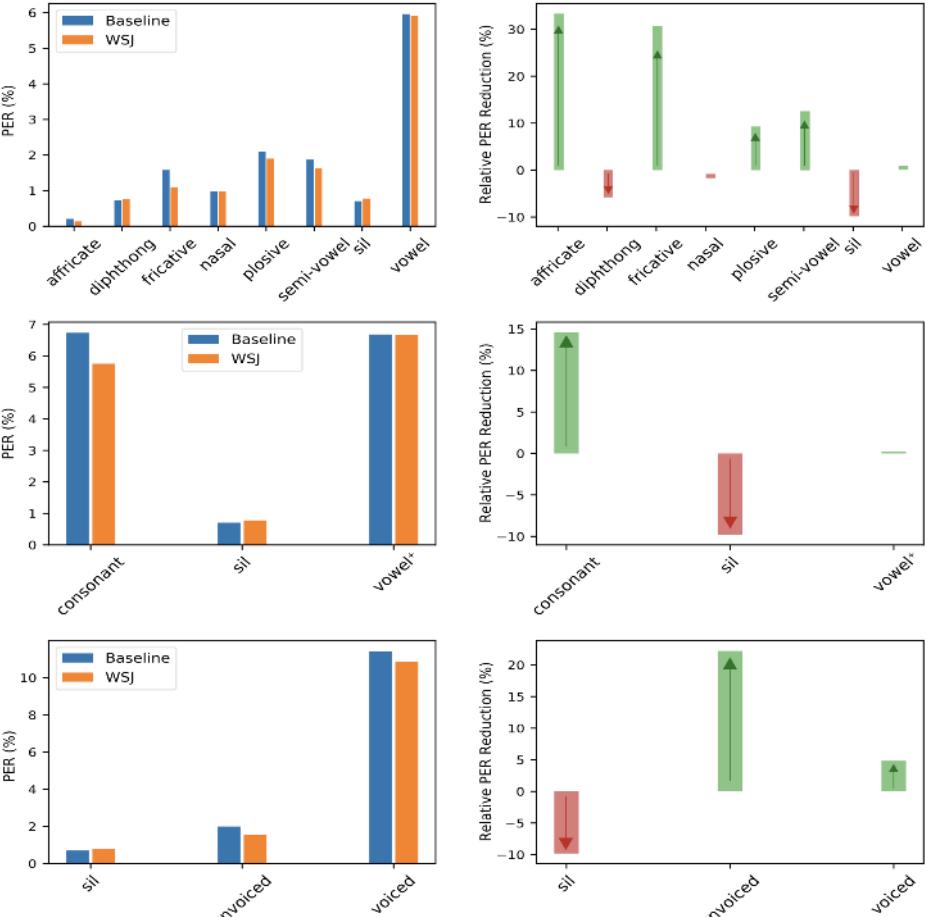
# Transfer Learning from WSJ

- Average relative gain:
  - Dev: 10%, Test: 7%



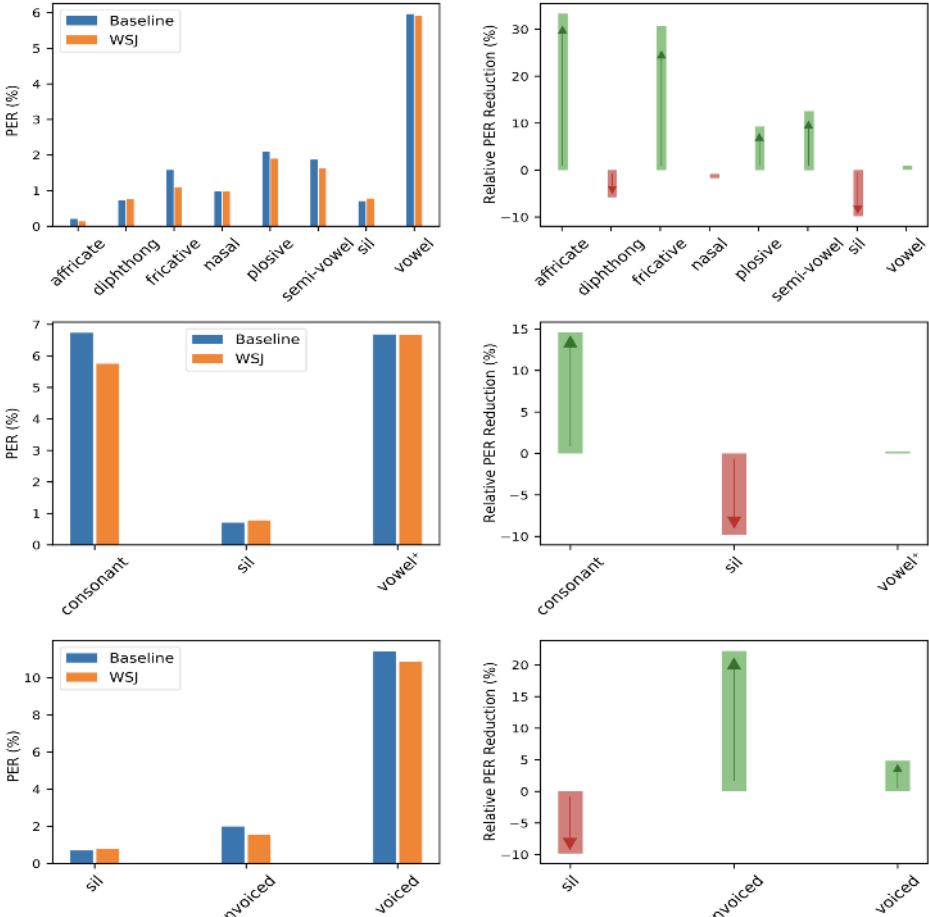
# Transfer Learning from WSJ

- Average relative gain:
  - Dev: 10%, Test: 7%
- Negative gain (**-10%**) for Silence!
  - Why?



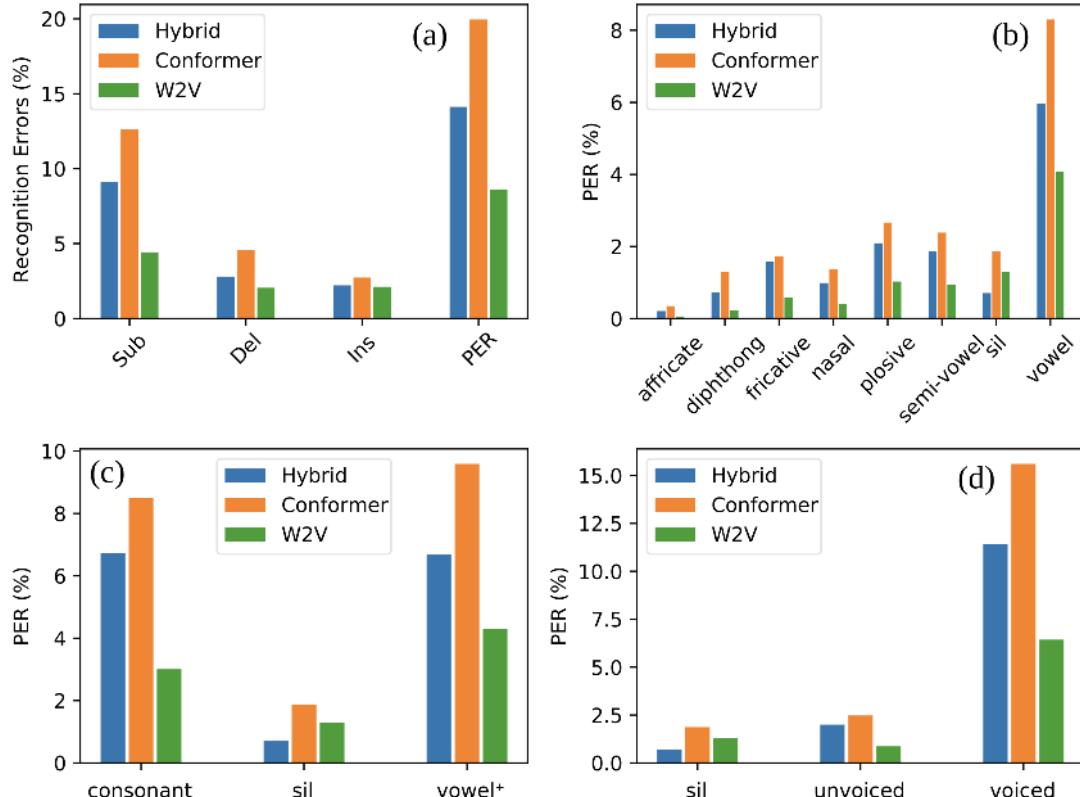
# Transfer Learning from WSJ

- Average relative gain:
  - Dev: 10%, Test: 7%
- Negative gain (-10%) for Silence!
  - Why?
- Average gain for  $C > V^+$ 
  - Why?

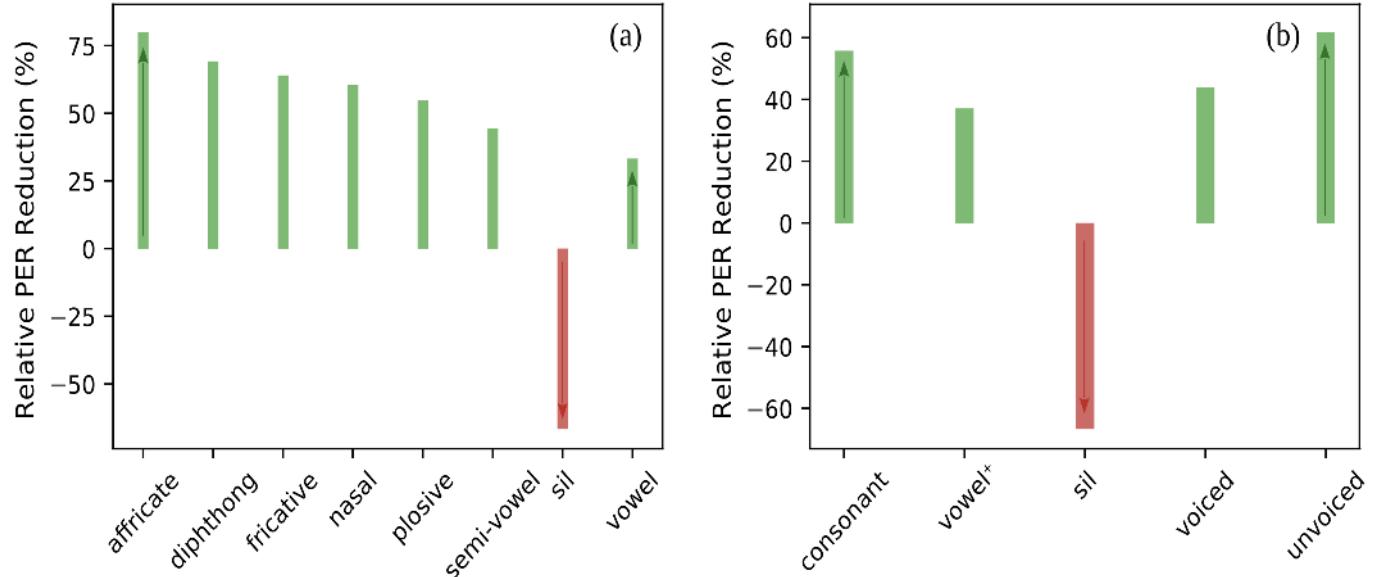


# End-to-end (E2E) vs Hybrid

- E2E systems
  - Conformer
  - Wav2vec 2.0
- Details in the paper ...



# HMM-DNN → Wav2Vec 2.0



- Relative Gain: +25% to 75%
- Negative gain (-60%) for Silence!
- Average gain C > V<sup>+</sup>

# Confusion Matrices

Baseline (DNN-HMM)

True Label

	aff	dip	fri	nas	plo	sem	sil	vow
aff	10	0	6	0	4	0	0	0
dip	0	13	0	1	1	13	0	50
fri	8	3	127	1	24	9	7	4
nas	0	1	3	41	9	4	3	5
plo	8	0	25	2	73	4	0	5
sem	5	16	12	3	7	18	2	54
sil	0	0	4	4	3	2	0	1
vow	1	48	4	5	5	48	3	549

Legend

aff: affricate  
dip: diphthong  
fri: fricative  
nas: nasal  
plo: plosive  
sem: semi-vowel  
sil: silence  
vow: vowel  
con: consonant  
sil: silence  
vow<sup>+</sup>: vow+dip  
sil: silence  
unv: unvoiced  
voi: voiced

Predicted Label

(a)

	sil	con	vow <sup>+</sup>
sil	0	13	1
con	12	403	88
vow <sup>+</sup>	3	78	660

(b)

	sil	unv	voi
sil	0	2	12
unv	5	55	84
voi	10	125	965

Wav2vec 2.0

True Label

	aff	dip	fri	nas	plo	sem	sil	vow
aff	0	0	0	0	2	1	0	0
dip	0	0	0	0	0	1	1	8
fri	1	0	37	0	2	1	2	0
nas	0	0	0	12	0	0	2	0
plo	2	0	2	0	19	0	1	0
sem	1	0	1	0	1	1	0	28
sil	0	0	1	0	8	1	0	0
vow	0	19	2	2	0	16	1	373

Predicted Label

(a)

	sil	con	vow <sup>+</sup>
sil	0	10	0
con	5	83	28
vow <sup>+</sup>	2	21	400

(b)

	sil	unv	voi
sil	0	1	9
unv	1	4	27
voi	6	24	477

# Confusion Matrices

Baseline (DNN-HMM)

True Label

	aff	dip	fri	nas	plo	sem	sil	vow
aff	10	0	6	0	4	0	0	0
dip	0	13	0	1	1	13	0	50
fri	8	3	127	1	24	9	7	4
nas	0	1	3	41	9	4	3	5
plo	8	0	25	2	73	4	0	5
sem	5	16	12	3	7	18	2	54
sil	0	0	4	4	3	2	0	1
vow	1	48	4	5	5	48	3	549

Legend

aff: affricate  
dip: diphthong  
fri: fricative  
nas: nasal  
plo: plosive  
sem: semi-vowel  
sil: silence  
vow: vowel  
  
con: consonant  
sil: silence  
vow<sup>+</sup>: vow+dip  
  
sil: silence  
unv: unvoiced  
voi: voiced

Predicted Label

(a)

Sparser but Similar Patterns

Wav2vec 2.0

True Label

	aff	dip	fri	nas	plo	sem	sil	vow
aff	0	0	0	0	2	1	0	0
dip	0	0	0	0	0	1	1	8
fri	1	0	37	0	2	1	2	0
nas	0	0	0	12	0	0	2	0
plo	2	0	2	0	19	0	1	0
sem	1	0	1	0	1	1	0	28
sil	0	0	1	0	8	1	0	0
vow	0	19	2	2	0	16	1	373

Predicted Label

Legend

aff: affricate  
dip: diphthong  
fri: fricative  
nas: nasal  
plo: plosive  
sem: semi-vowel  
sil: silence  
vow: vowel  
  
con: consonant  
sil: silence  
vow<sup>+</sup>: vow+dip  
  
sil: silence  
unv: unvoiced  
voi: voiced

(b)

	sil	con	vow <sup>+</sup>
sil	0	13	1
con	12	403	88
vow <sup>+</sup>	3	78	660

(c)

	sil	unv	voi
sil	0	2	12
unv	5	55	84
voi	10	125	965

(b)

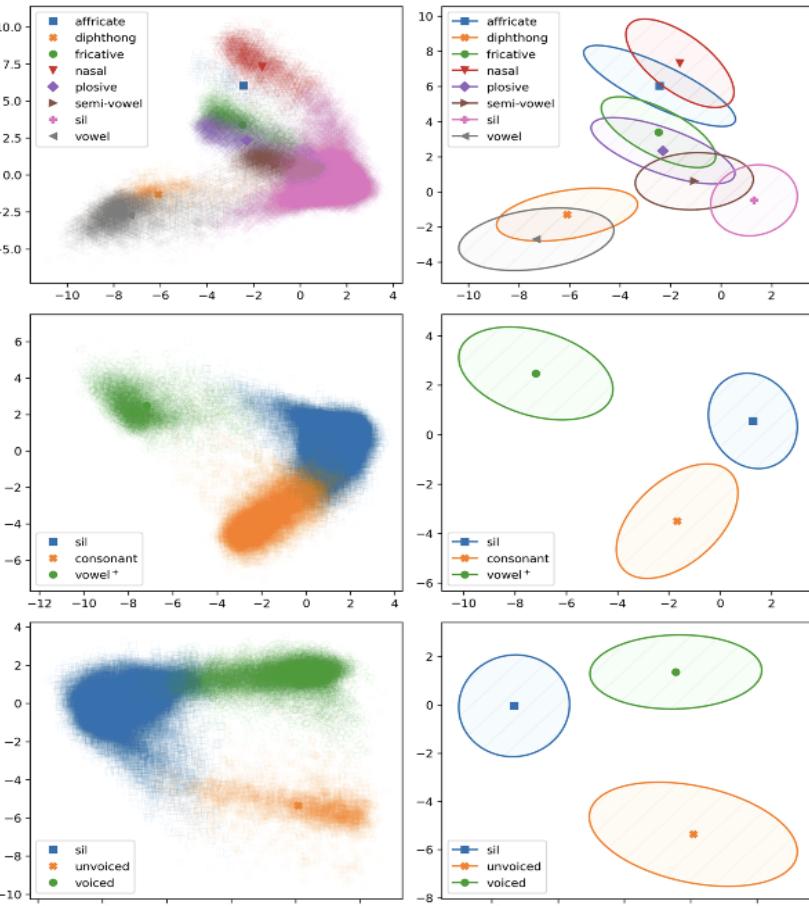
	sil	con	vow <sup>+</sup>
sil	0	10	0
con	5	83	28
vow <sup>+</sup>	2	21	400

(c)

	sil	unv	voi
sil	0	1	9
unv	1	4	27
voi	6	24	477

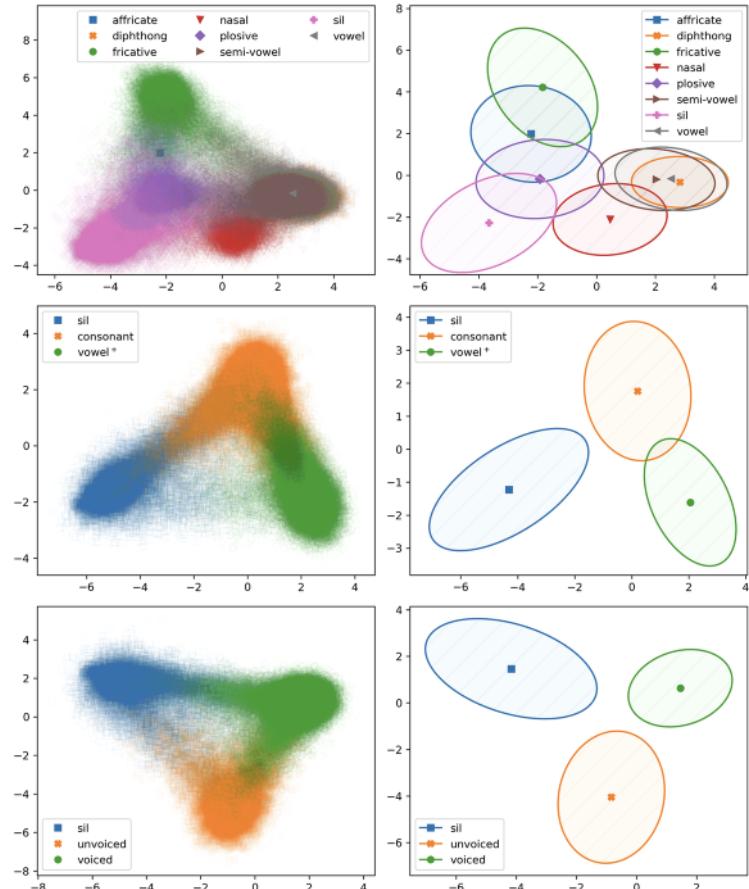
# Scatter Plots

@Logit (wav2vec 2.0)



# Scatter Plots

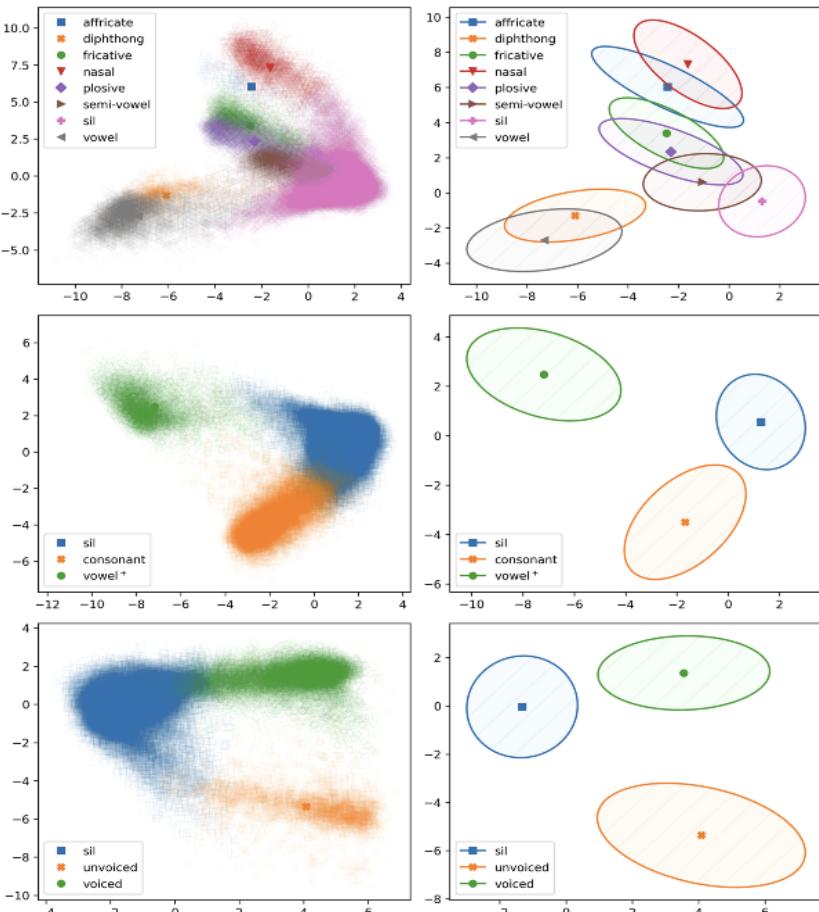
@Logit (Baseline) ↔ Hybrid)



More distinct clusters

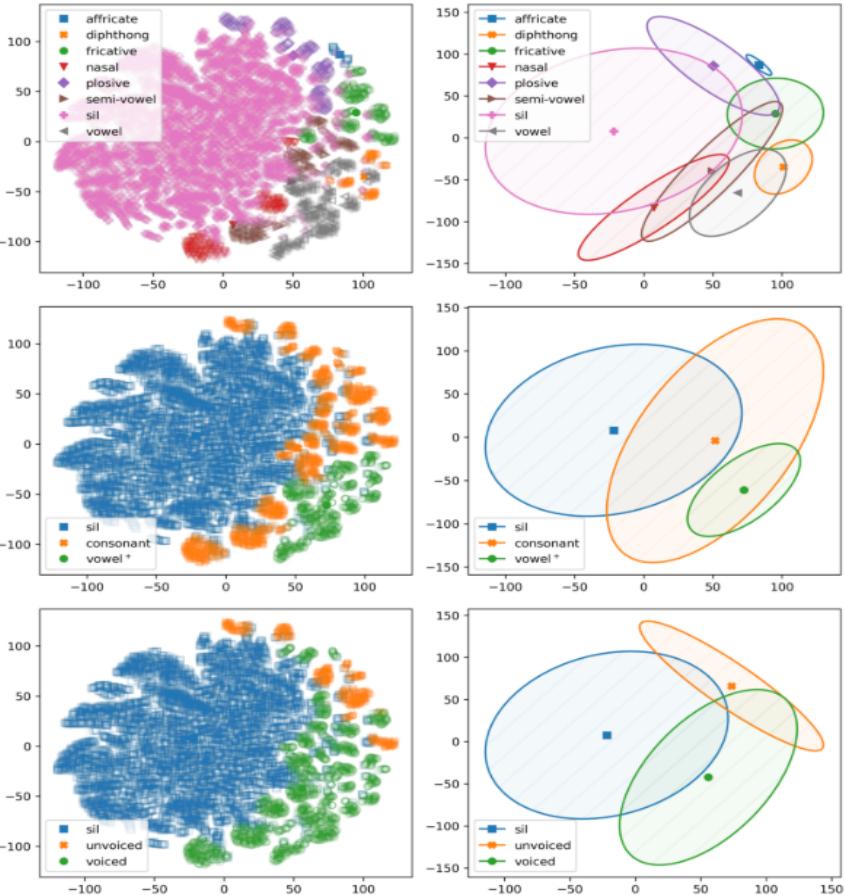


@Logit (wav2vec 2.0)



# t-SNE vs LDA

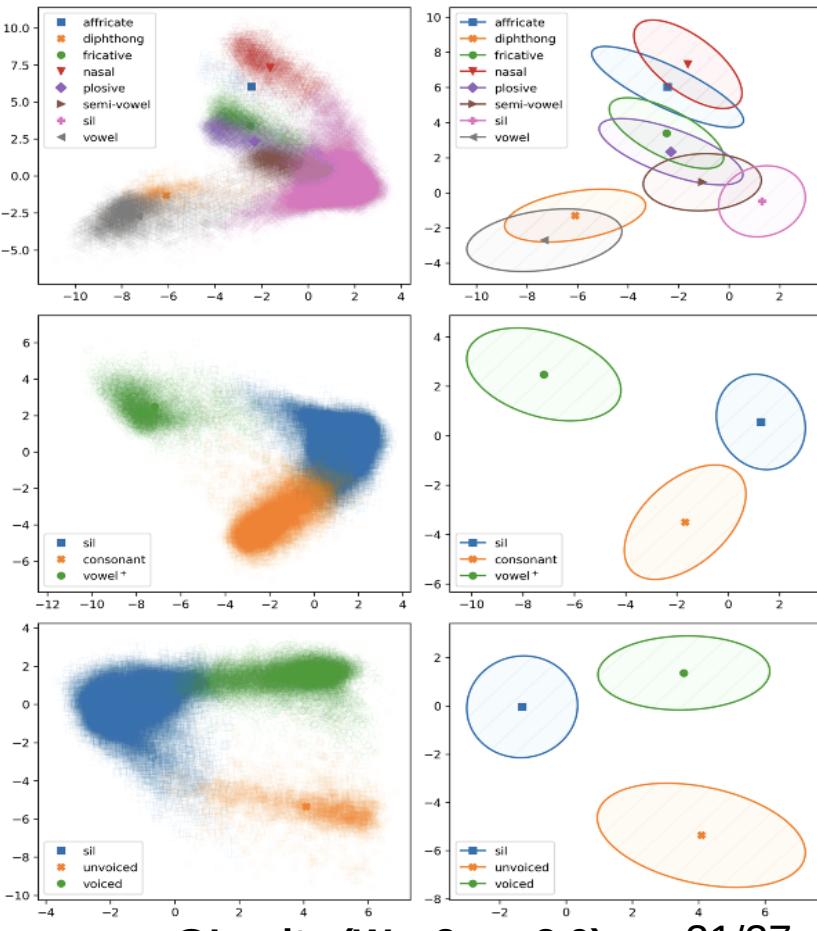
t-SNE



@Logits (Wav2vec 2.0)

Loweimi et al

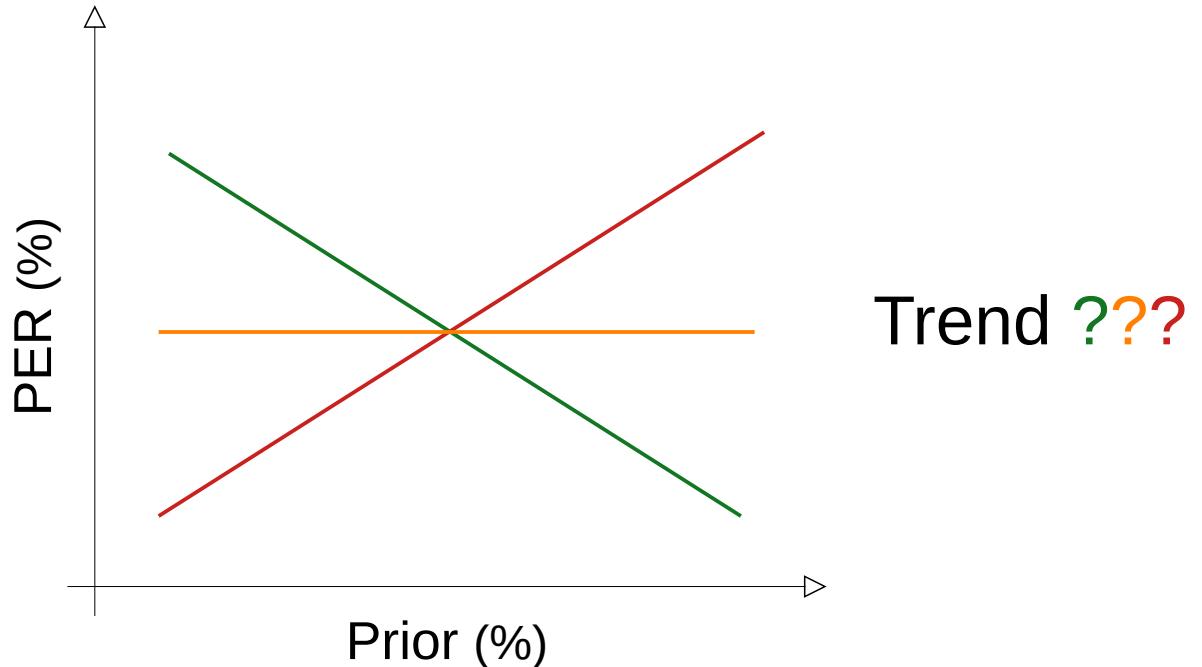
LDA



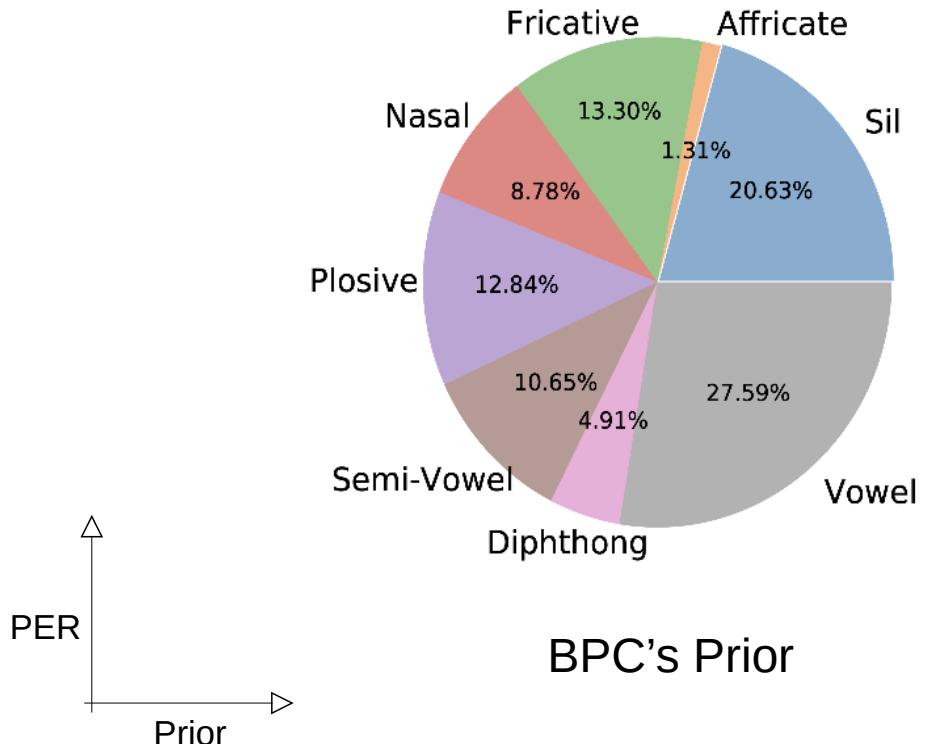
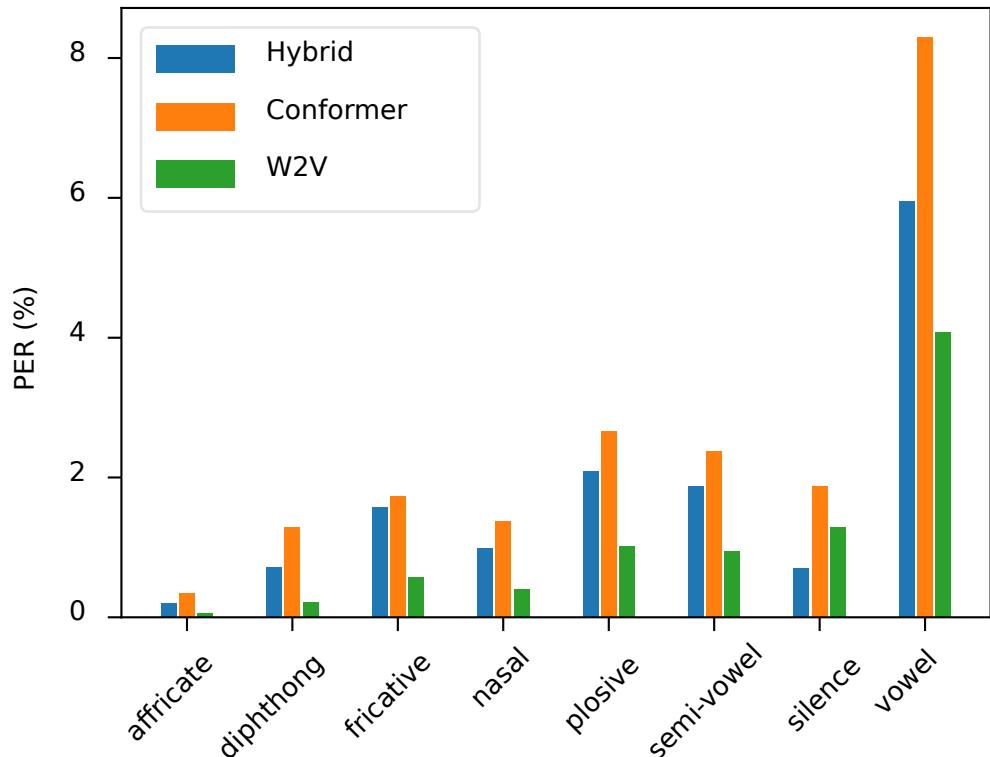
@Logits (Wav2vec 2.0)

31/37

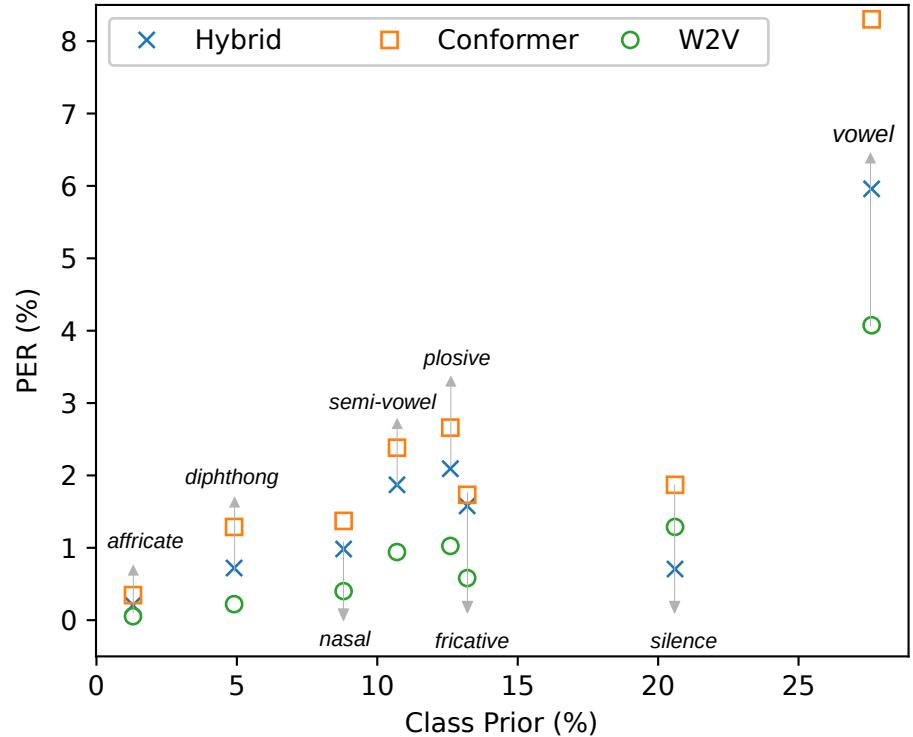
# PER vs Prior (1)



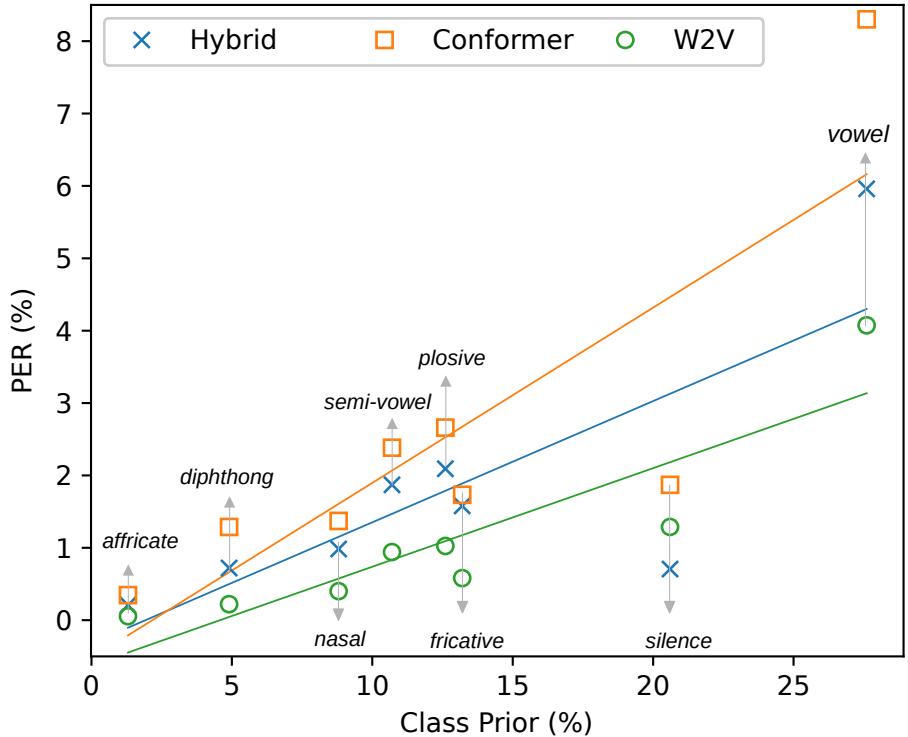
# PER vs Prior (1)



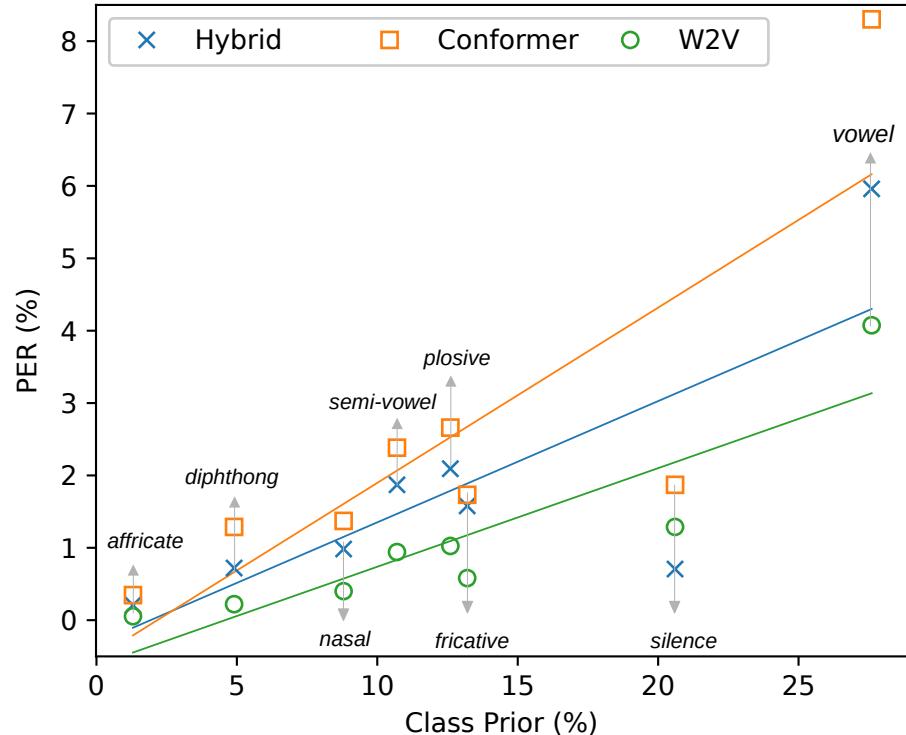
# PER vs Prior (2)



# PER vs Prior (2)

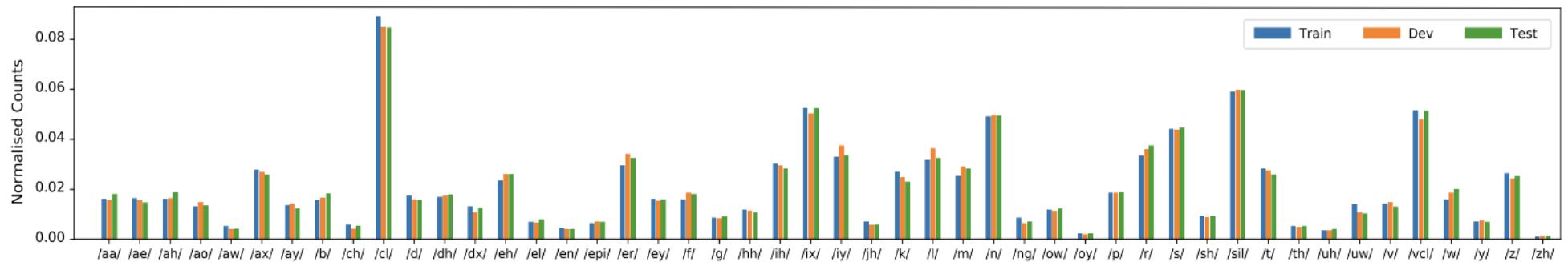


# PER vs Prior (2)

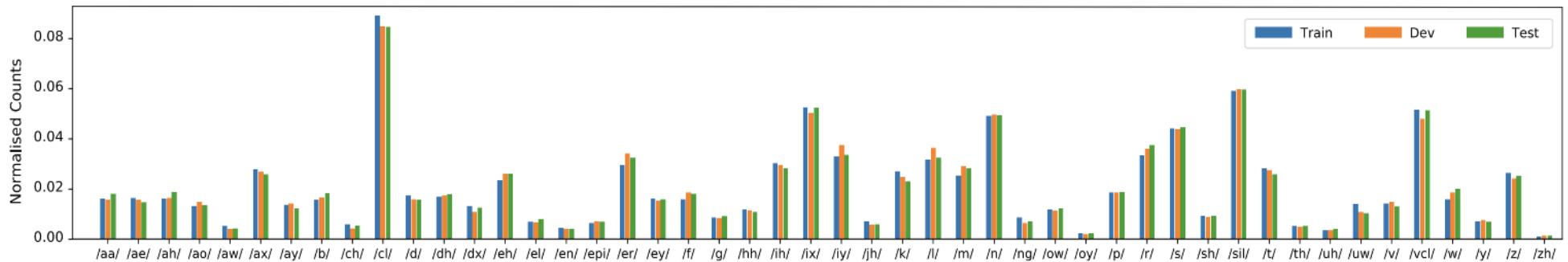


The higher the training data, the higher the PER!

# TIMIT's Special Case ...



# TIMIT's Special Case ...



The higher the training data, the larger/richer the test set, the higher the PER!

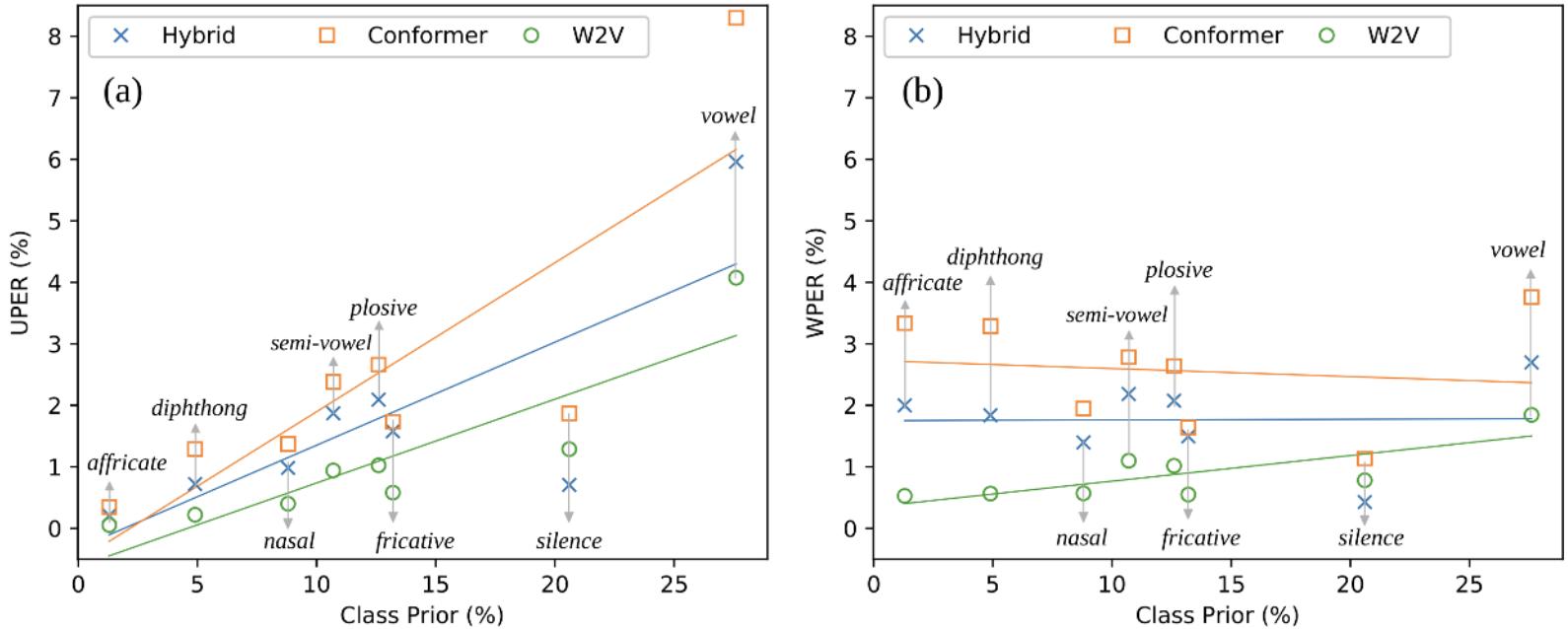
# Unweighted vs Weighted PER

$$PER = UPER = \frac{1}{N} \sum_{c=1}^C (Sub_c + Del_c + Ins_c)$$

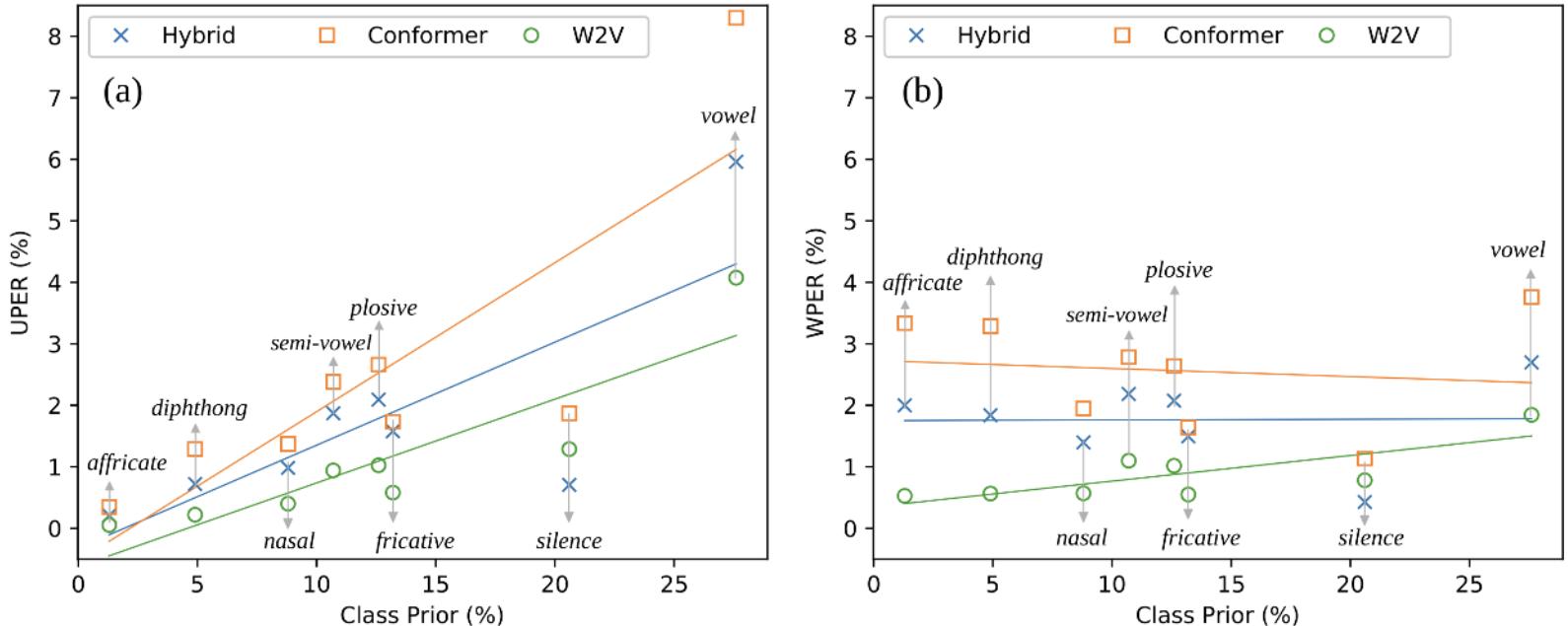
$$WPER = \frac{1}{C} \sum_{c=1}^C \frac{Sub_c + Del_c + Ins_c}{N P_c}$$

- To compensate for non-uniform prior ( $P_c \neq 1/C$ ), we use WPER.
- Analogous to *weighted accuracy*, weight  $\propto 1/P_c$ .

# UPER & WPER vs Prior



# UPER & WPER vs Prior



- Weighting flattens PER vs Prior
- Vowels still have the largest PER

# Wrap-up

- **Goal:** Break down PER, using broad phonetic classes
- **Findings:**
  - Largest PER share belongs to Vowels
  - Training dynamics is similar for all, except Silence
  - Uni → bi-directional seq. modelling is least useful for Silence
  - GMM → DNN is most useful for Silence
  - **Most/Least** robust classes to noise (NTIMIT) → **Vowels/Fricatives**
  - Transfer learning or pre-training are useful for all, except Silence
  - Consonants benefit more than Vowels from more data

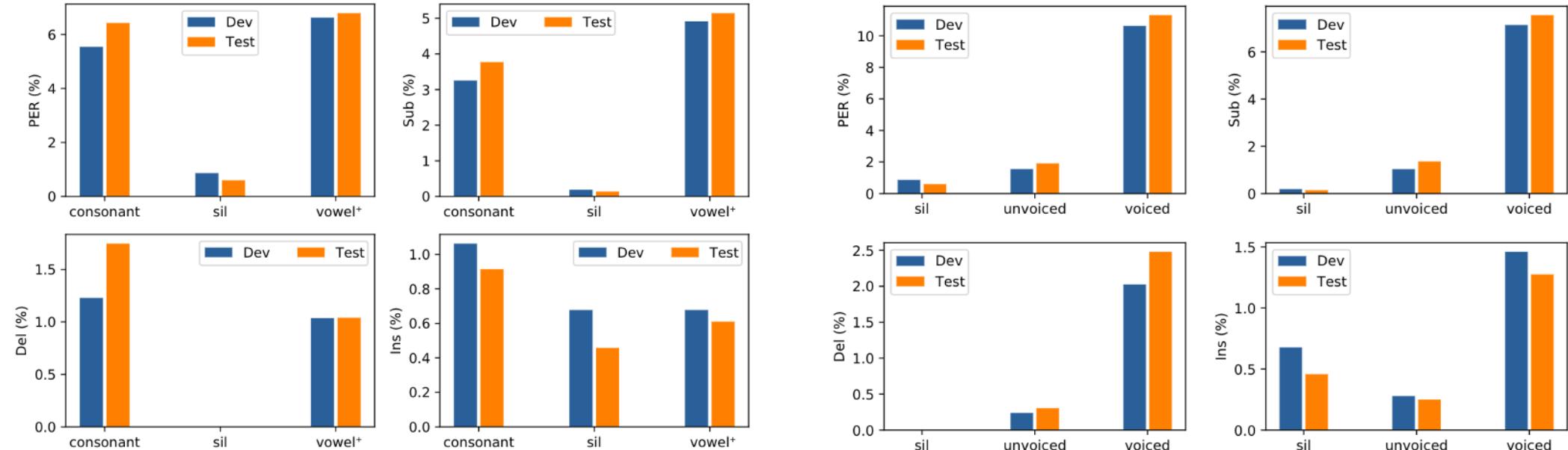
# That's It!

- Thank you!
- Q&A
- Appendices
  - (A1) Silence Class
  - (A2) Sub/Del/Ins of C/V<sup>+</sup> & V/U
  - (A3) Sub/Del/Ins Dynamics
  - (A4) Transfer Learning's Effect on Errors & Dynamics

# (A1) Silence Class is Union of ...

- /h#/ : Silence ( $\equiv$  non-speech) at Beginning & End
- /epi/ : epenthetic silence between a **fricative** & a **nasal** or **semi-vowel**
  - Within a word: **small**, **prince**
  - between words: “... *lines must* ...” (sx352.phn)
- /pau/ : short pause within a sentence
- /closures/ : closures before stops (plosives)
  - Union of /bcl/, /dcl/, /gcl/, /kcl/, /pcl/ and /tcl/

# (A2) Sub/Del/Ins Errors for C/V<sup>+</sup> & V/U



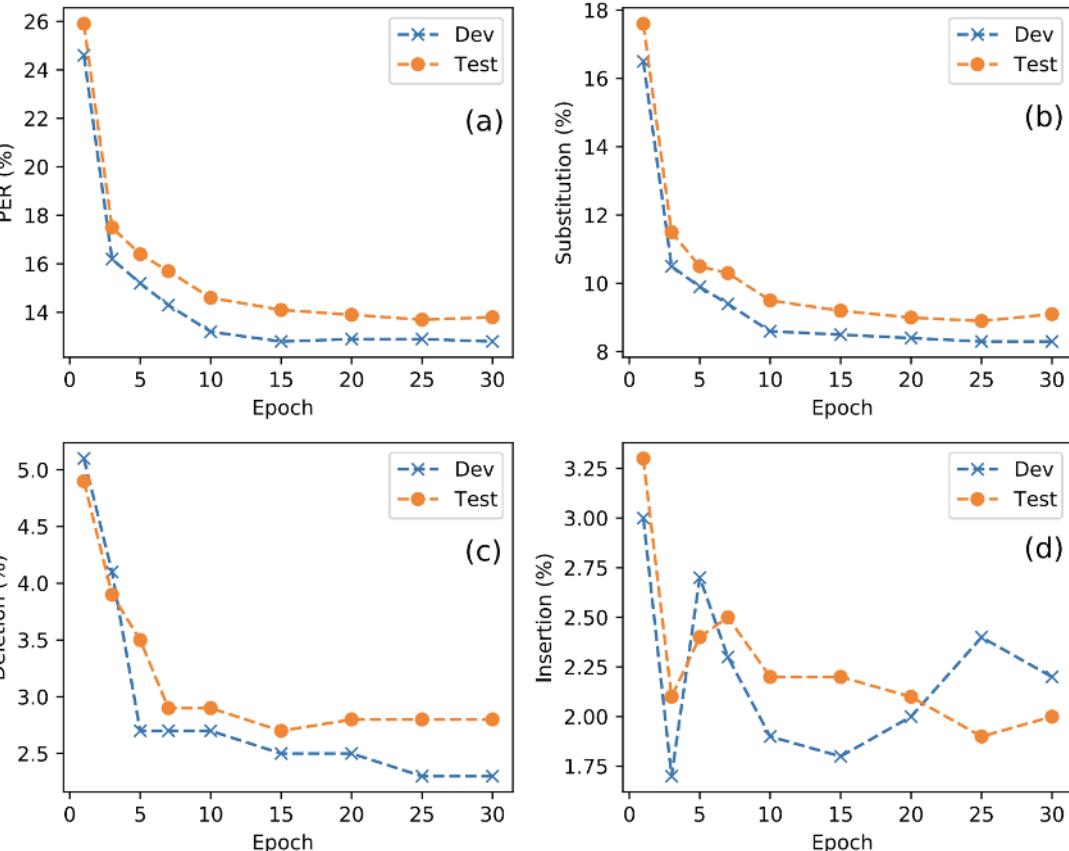
- Vowels have larger Sub.
- Consonants have a larger Del & Ins.

Errors' Largest share belongs to Voiced.

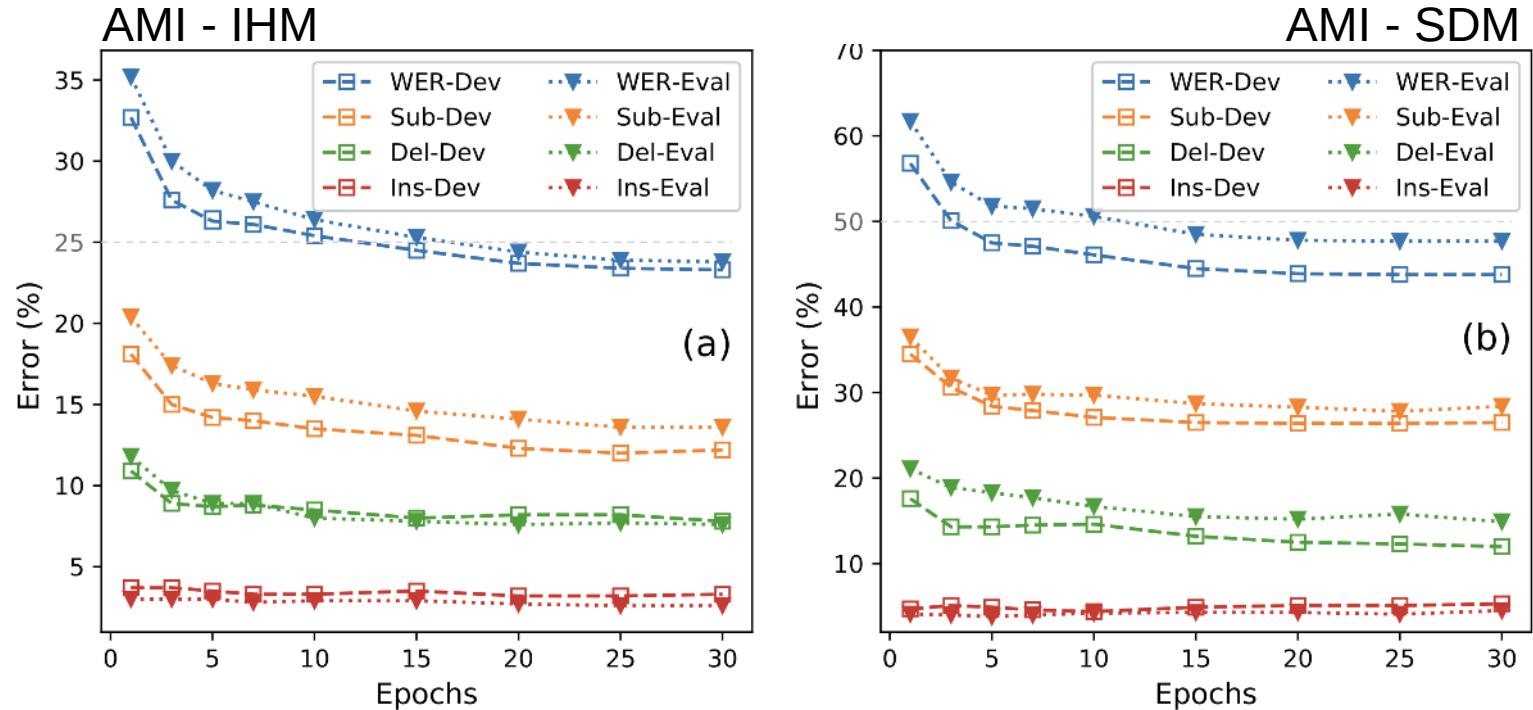
# (A3) Sub/Del/Ins Dynamics

Strongest Correlation: PER & Sub

- Sub converges slowly
- Del converges fast
- Ins oscillates



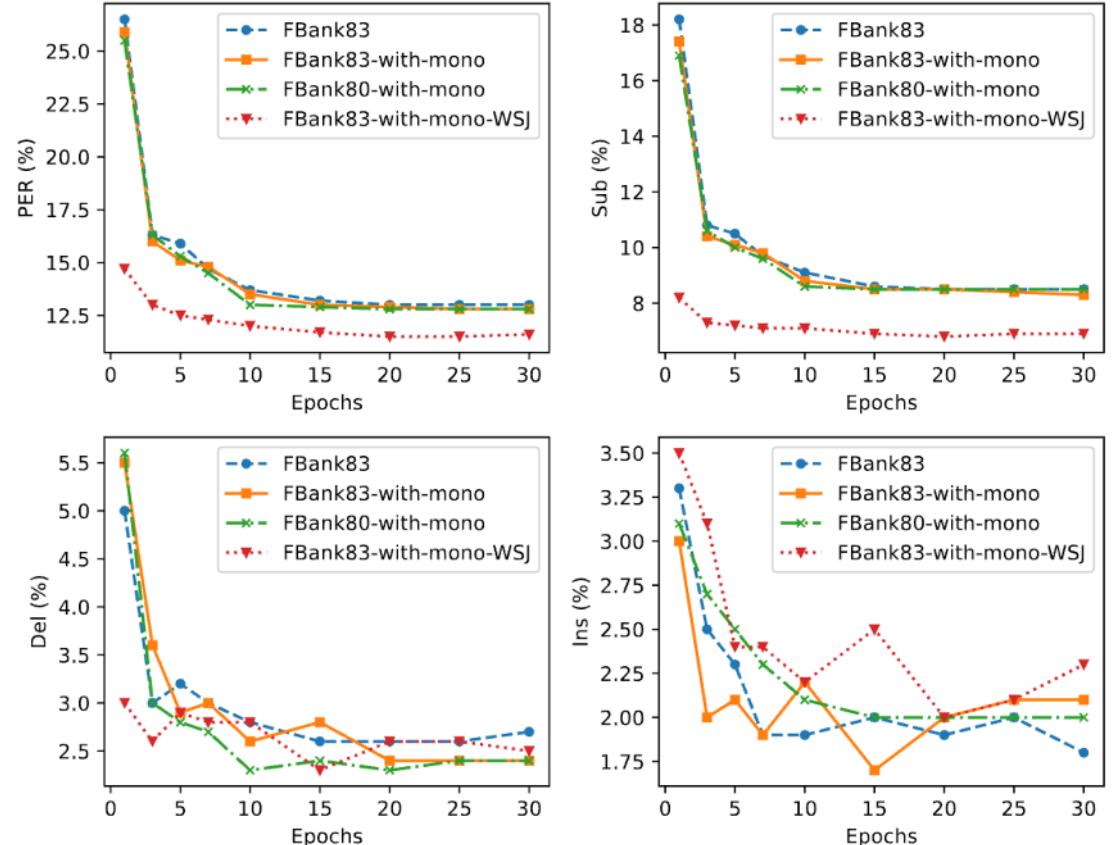
# (A3) Similar observations in [67]



(Fig. 15 in) E. Loweimi, Z. Yue, P. Bell, S. Renals, and Z. Cvetkovic, "Multi-stream Acoustic Modelling using Raw Real and Imaginary Parts of the Fourier Transform", in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 876-890, 2023, doi: 10.1109/TASLP.2023.3237167.

# (A4) TF's Effect on Errors & Dynamics

- Dynamic range becomes smaller.
- Mostly improves Sub error.
- Slightly makes Ins worse.



\* TF: Transfer Learning

Loweimi et al

A3/4