Phonetic Error Analysis Beyond Phone Error Rate

Erfan Loweimi *(Member, IEEE)*, Andrea Carmantini *(Member, IEEE)*, Peter Bell *(Member, IEEE)*, Steve Renals *(Fellow, IEEE)*, Zoran Cvetkovic *(Senior Member, IEEE)*

***Abstract*—In this paper, we analyse the performance of the TIMIT-based phone recognition systems beyond the overall phone error rate (PER) metric. We consider three broad phonetic classes (BPCs): {affricate, diphthong, fricative, nasal, plosive, semi-vowel, vowel, silence}, {consonant, vowel, silence} and {voiced, unvoiced, silence} and, calculate the contribution of each phonetic class in terms of the substitution, deletion, insertion and PER. Furthermore, for each BPC we investigate the following: evolution of PER during training, effect of noise (NTIMIT), importance of different spectral subbands (1, 2, 4, and 8 kHz), usefulness of bidirectional vs unidirectional sequential modelling, transfer learning from WSJ and regularisation via monophones. In addition, we construct a confusion matrix for each BPC and analyse the confusions via dimensionality reduction to 2D at the input (acoustic features) and output (logits) levels of the acoustic model. We also compare the performance and confusion matrices of the BLSTM-based hybrid baseline system with those of the GMM-HMM based hybrid, Conformer and wav2vec 2.0 based end-to-end phone recognisers. Finally, the relationship of the unweighted and weighted PERs with the broad phonetic class priors is studied for both the hybrid and end-to-end systems.**

***Index Terms*—Phone recognition, TIMIT, phonetic error analysis, broad phonetic classes, confusion matrix, hybrid, end-to-end**

I. Introduction

THE performance of the phone recognition systems is commonly reported in terms of the phone error rate (PER) which is a minimum edit distance reflecting the *overall* number of substitution, deletion and insertion errors. While PER enables comparisons and rankings of phone recognisers, it lacks the granularity needed to understand the nature of errors and nuances in phone recognition.

The central research question this paper aims at exploring is as follows: what is the contribution of each *broad phonetic class* (BPC) to the overall PER? If from the acoustic phonetics perspective [1], we define the broad phonetic classes as *consonant*, *vowel* and *silence*, what proportion of PER is associated with each class? The broad phonetic classes could also be defined as *voiced*, *unvoiced* and *silence* or with a higher resolution as *affricate*, *diphthong*, *fricative*, *nasal*, *plosive* (*stop*), *semi-vowel*, *vowel* and *silence*.

The concept of broad phonetic classes has been explicitly and implicitly employed in a wide range of applications. Yuan and Liberman [2] used broad phonetic classes for speaking rate and syllable detection and demonstrated such systems are more robust than monophone based ones. Ludusan and Dupoux [3] utilised BPCs for syllable segmentation based on the sonority sequencing principle [4]. In [5] and [6], BPCs were used for speaker verification and identification and both observed that in these tasks the vowels and nasals are more useful than other phonetic classes. Lu *et al.* [7] leveraged broad phonetic class posteriorgram in the speech enhancement task, showcasing that they can contribute towards enhancing both speech quality and intelligibility. BPCs have also been used in the time-scale modification [8], [9] to improve the perceptual quality. They also have found application in speech coding [10] in order to allocate different number of bits to speech frames. For example, the source-controlled variable rate coder proposed in [11] operates with rates of 4.9, 3.0 and 0.67 kbps for voiced, unvoiced and silence sounds, respectively. Kempton and Moore illustrated the usefulness of BPCs in the language identification task [12]. Lee *et al.* [13], Ringval *et al.* [14] and Yuan *et al.* [15] explored the use of the broad phonetic classes in speech emotion recognition task and demonstrated that vowels are the most useful phonetic class for speech emotion recognition. Additionally, BPCs were applied for phone recognition as a training criterion in [16], [17], [18], [19] and also to develop noise-robust segment-based phone recognisers [20]. Young *et al.* used the broad phonetic class concept for constructing decision trees [21] which are widely employed in the hybrid speech recognition systems. Sainath [22] applied BPCs in detecting *islands* (reliable speech segments) in order to prune the search space for speech recognition. Gravier *et al.* [23] and Ziegler *et al.* [24] proposed landmark-based and frame-based approaches, respectively, for detecting BPCs to guide the decoding process in ASR. The broad phonetic classes were also leveraged in multi-lingual ASR [25], [26] and spectrogram reading [27], as demonstrated in [28]. Finally, the BPCs are implicitly utilised by infants during the phonological development and language acquisition [29], [30], [31]. For example, Hochmann *et al.* [30] demonstrated that 12-month-old infants rely more on consonants when identifying words, while they are better at recognising and generalising patterns that are based on repetition of vowels.

This paper aims at using the concept of broad phonetic classes for error analysis on the TIMIT [32] phone recognition task. That is, we decompose the overall PER into the contribution of each category within the three broad phonetic groups defined earlier. Furthermore, we study the confusion patterns

Manuscript received 30 May 2023; revised 4 August 2023; accepted 5 September 2023. Date of current version 5 September 2023. This work was supported by EPSRC Project under Grant EP/R012180/1 (SpeechWave). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Samuel Thomas.(Corresponding author: Erfan Loweimi.)

E. Loweimi (corresponding author) and Z. Cvetkovic are with King's College London (KCL); e-mail: {erfan.loweimi, zoran.cvetkovic}@kcl.ac.uk. A. Carmantini, P. Bell and S. Renals are with the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK; e-mail: {a.carmantini, peter.bell, s.renals}@ed.ac.uk.

Supported by EPSRC Project EP/R012180/1 (SpeechWave).

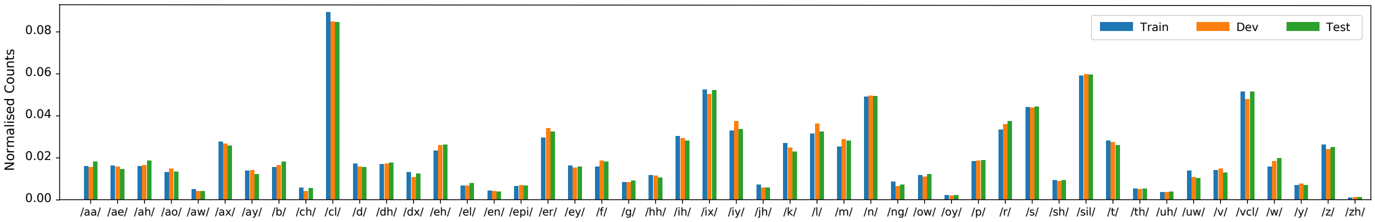


Fig. 1. Normalised count (class prior probability) for TIMIT's standard train, dev and test sets over 48 phones. Counts of the phones in the train, dev and test data are normalised by the number of frames for each subset, namely 1124823, 122487 and 57919, respectively (frame length: 25 ms, frame shift: 10 ms).

within these broad classes by constructing three confusion matrices. To investigate and analyse the observed confusions, we deploy the scatter plots of the acoustic model's input (i.e., acoustic feature) and output (i.e., logits) after dimensionality reduction to 2D via linear discriminant analysis (LDA) [33].

Moreover, for each broad phonetic class, we study the training dynamics in terms of PER vs epoch, investigate importance of different spectral subbands, usefulness of bidirectional vs unidirectional sequential modelling, transfer learning from WSJ [34] and effect of noise (NTIMIT [35]). Finally, we compare the performance and confusion matrices of the state-of-the-art end-to-end (Conformer [36] and wav2vec 2.0 [37]) and hybrid phone recognition systems.

The rest of this paper is structured as follows. In Section 2, we provide a review of the TIMIT database and define the three broad phonetic classes used in our analysis. Section 3 describes the experimental setup and presents initial results. Section 4 focuses on the phonetic error analysis of the baseline system. In Section 5, we compare various modeling factors and systems with the baseline. Finally, Section 6 concludes the paper and highlights potential avenues for future research.

TABLE I
MAPPING TO THE 8-CLASS BROAD PHONETIC CLASSES.

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

TABLE II
MAPPING TO THE CONSONANT, VOWEL$^+$, SILENCE, VOICED AND UNVOICED BPCS.

classes	phones
Vowel$^+$	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

II. PHONE RECOGNITION BY TIMIT

A. *TIMIT database*

TIMIT [32] has been widely used as a benchmark for acoustic-phonetics studies (e.g., [38], [39]), phone recognition (e.g., [40], [41], [42]), phone classification (e.g., [43], [44]), phone segmentation (e.g., [45], [46]) and speaker recognition (e.g., [47], [48]). It consists of 5.4 hours of speech, manually transcribed at the word and phone levels. Although the amount of data is not favourably large towards building large-scale deep neural networks (DNNs), it was among the tasks used to verify the effectiveness of DNNs [49], [50] and is still widely applied in evaluating various models and ideas (e.g., [37]).

TIMIT contains sentences read by 630 speakers (192 female and 438 male) of eight major dialects of American English, each reading ten phonetically rich sentences. There are three types of sentences: SA (to express speakers' dialectal variances), SX (phonetically compact) and SI (phonetically diverse). There are 2 SA, 450 SX and 1890 SI distinct sentences and each talker reads 2 SA, 5 SX and 3 SI sentences. The SA ones are read by all speakers, each SX sentence is read by 7 speakers and each SI sentence is read by a single speaker.

TIMIT contains 5.4 hours of speech, 5107 unique words, with 8.2 words per sentence on average [51]. The standard train set comprises of 3.14 hours data (2310 SX and 1386 SI sentences). The dev set consists of 21 minutes of speech (250 SX and 150 SI sentences). The test set includes 10 minutes of speech (120 SX and 72 SI sentences). SA sentences are not included in these sets.

For each utterance, four files are provided: the speech waveform (16 kHz sampling rate, 16-bit resolution), orthographic transcription, time-aligned phonetic transcription and word level transcription. The original phonetic alignments in TIMIT include 61 phones. In [52], some phonetic classes have been merged leading to two sets of 48 and 39 phones, often employed for training and evaluation purposes, respectively. To perform such a mapping from 61 to 48 to 39, we utilise `phones.60-48-39.map` [53]. Note that although TIMIT's original transcription includes 61 phones, in naming this file 60 is used. This is owing to excluding the *stop closure* /q/.

Fig. 1 shows the amount of data, in terms of number of frames per phone, after mapping to the 48-phone set. We normalised the number of frames per phone by the total number of frames to get normalised counts reflecting the class prior probabilities. As seen, a priori probabilities per phonetic class are almost identical across the standard train, dev and test sets, demonstrating the careful design of the data set.

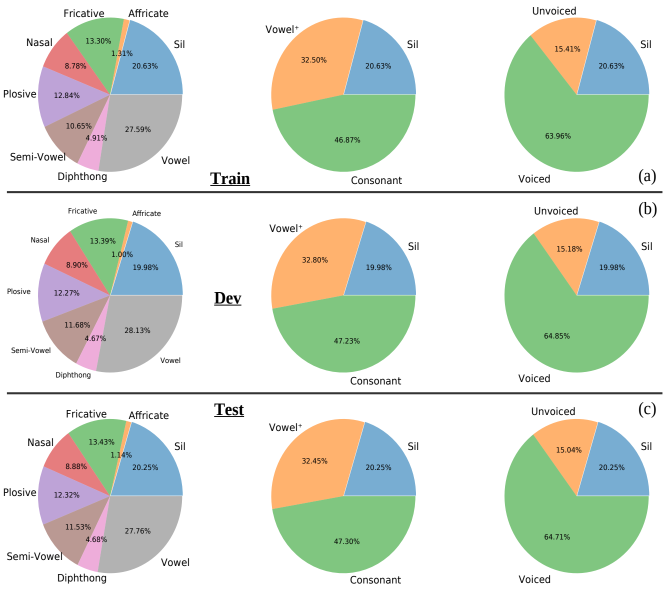


Fig. 2. BPCs' prior probabilities (%) for TIMIT's train, dev and test sets.

B. Broad phonetic classes

In this paper, we consider three broad phonetic categories:

1) affricate, diphthong, fricative, nasal, plosive, semi-vowel, vowel and silence, as defined in Table I;
2) consonant, vowel$^+$ and silence, as defined in Table II;
3) voiced, unvoiced and silence, as defined in Table II.

We refer to the first category as *8-class*. Note that

- the phones are first mapped to the 39-phone set and then to broad phonetic classes (BPC) via Tables I and II;
- by silence we mean silence at the beginning/end (/h#/), pauses (/pau/), epenthetic silence (/epi/) and closures; the closures in TIMIT's original phonetic transcription are as follows: /bcl/, /dcl/, /gcl/, /kcl/, /pcl/ and /tcl/;
- semi-vowels are difficult to characterised as are produced like vowels but function as consonants; based on (Table 2.8 in) [54], we categorise them as consonants;
- vowel$^+$ class is the union of vowels and diphthongs;
- silence is identical across all three BPC definitions.

Fig. 2 illustrates the priors for the three broad phonetic classes over TIMIT's train, dev and test sets. As seen, TIMIT data is greatly balanced over all sets. For example, the prior for the fricative, consonant, voiced and unvoiced phones across the train/dev/test data are 16.3/16.3/16.5%, 35.3/35.4/34.5%, 63.3/64.1/63.1% and 17.3/16.9/17.8%, respectively.

Fig. 2 also shows the distribution across different broad phonetic classes is not uniform. For example, 63.3% of training data is voiced, 17.3% is unvoiced, 19.4% is silence, 35.3% is vowel$^+$ and 45.3% is consonant. In addition, 1.4% is affricate, 16.3% is fricative, 7.9% is nasal, 9.4% is plosive, 10.3% is semi-vowel, 8% is diphthong and 27.3% is vowel. From this perspective, TIMIT is not a balanced dataset. This property, however, is a well-known characteristic of natural languages and explored by several linguistic theories such as the Quantal Theory [55], [56] and the Theory of Adaptive Dispersion [57].

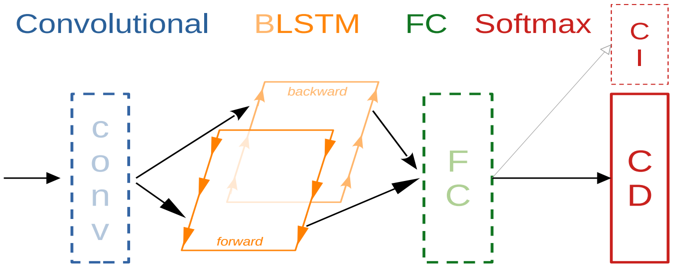


Fig. 3. The acoustic model of the hybrid system consists of bidirectional LSTM (BLSTM) layers, possibly preceded by convolutional (conv) and succeeded by fully-connected (FC) layers. The output layer composed of the context-dependent (CD) and possibly context-independent (CI) heads.

TABLE III
PERS OF VARIOUS ACOUSTIC MODELS ON TIMIT
(*: WITHOUT REGULARISATION BY CI).

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	**12.8**	**14.1**	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

III. EXPERIMENTAL SETUP

A. Choosing acoustic model

We wish to build a phone recogniser with the state-of-the-art performance on TIMIT. We start with hybrid systems and later compare the results with the end-to-end (E2E) models.

To construct a hybrid systems we consider a wide range of acoustic models consisting of bidirectional long short-term memory (BLSTM) [58] layers along with possibly convolutional and fully-connected (FC) layers, as shown in Fig. 3. The best system is selected for the phonetic error analysis.

B. Setup

We use 83-D filterbank (FBank-83) features (80 filters along with three pitch-related representations [59]). Features are mean-variance normalised at the speaker level. In Table III, *CiLjFk* denotes a cascade of i convolutional layers, j BLSTM layers and k fully-connected layers followed by a softmax output layer. When number of layers for a specific layer type is zero, that layer is removed from the architecture name. For example, *L4* means an acoustic model consisting of only four BLSTM layers (as well as a softmax layer).

If included, the fully-connected layer contains 1024 nodes and the convolutional layer consists of 80 kernels of length 10 with a max pooling of size 3. Dropout [60] and ReLU [61] activation function are used in both convolutional and FC layers. BLSTM layers contain 550 nodes in each direction along with dropout. Batch normalisation [62] was used in both BLSTM and FC layers, and the batch size was set to 8. DNNs

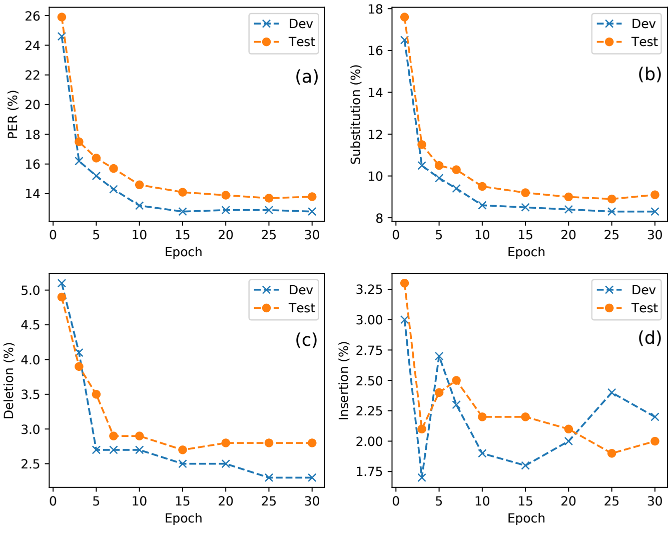


Fig. 4. Evolution of PER, substitution, deletion and insertion errors vs epoch.

were trained by the PyTorch-Kaldi toolkit [63], [64], [53] with RMSprop [65] optimiser. The models are trained by the cross entropy (CE) loss computed using context-dependent (CD) state-clustered triphones as well as context-independent (CI) monophones (for regularisation purposes [66]). The CD and CI output heads consist of 1936 and 48 nodes, respectively.

C. Initial results and discussion

Table III shows the highest performance is achieved by the system with four BLSTM layers (L4). The L4 system leads to 12.8% and 14.1% PER for TIMIT's dev and test data, respectively. This is a competitive performance on TIMIT when only the original training data is used. Additionally, incorporating pitch-related features and regularisation with monophones have a minor positive effect on the performance.

Fig. 4 depicts the performance evolution in terms of the PER, substitution (Sub), deletion (Del) and insertion (Ins) errors. The Sub, Del and Ins errors are relatively responsible for about 65%, 20% and 15% of PER, respectively. The Sub error is the dominant component of PER and its dynamics highly resembles that of PER. The Del error converges faster, while the Ins error oscillates during training. A similar observation was reported in (Fig. 15 of) [67], where evolution of the word error rate (WER), Sub, Del, and Ins errors vs epoch were analysed for the ASR task on the AMI meeting corpus [68].

When comparing the characteristics of Del and Ins errors with Sub errors, it becomes evident that minimising Sub errors is more intricate and demands additional epochs. This complexity arises from the need to transform one character into another, encompassing a broader range of possibilities and potential ambiguities. Consequently, achieving accurate correction necessitates a deeper understanding of the underlying context. In contrast, Del and Ins errors are relatively simpler to manage as they involve either removing or adding tokens, requiring less complex decision-making for correction.

These metrics reflect the overall performance. Having chosen the acoustic model, in the next section we present and discuss errors made by each broad phonetic class.

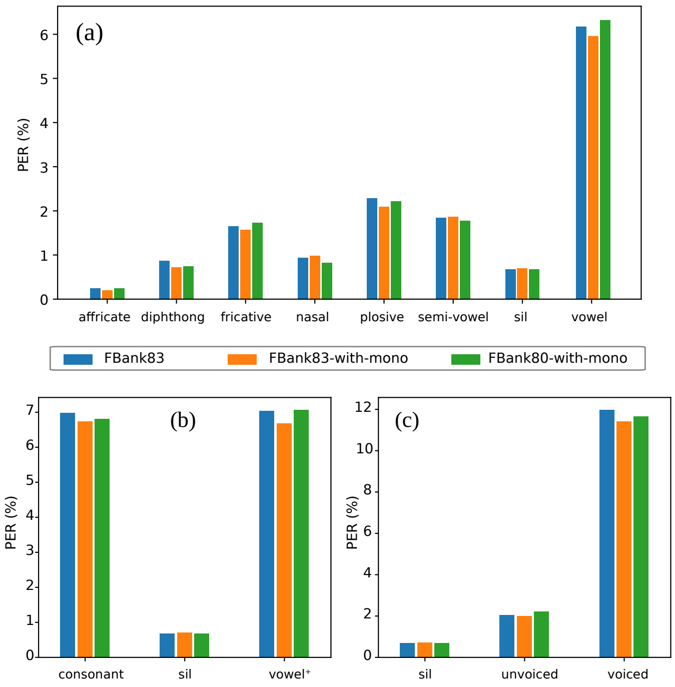


Fig. 5. PER for different broad phonetic classes: (a) 8-class, (b) consonant/vowel$^+$/silence, (c) voiced/unvoiced/silence.

IV. Phonetic Error Analysis

To evaluate PER for each broad phonetic class, we first map each phone to the corresponding broad categories using Tables I and II. Then, the substitution, deletion and insertion errors are accumulated and finally, the PER per class is reported.

A. Comparison of acoustic models

We start by comparing PER of different broad phonetic classes for three acoustic models from Table III: FBank-83 (without regularisation by monophones (CI)) as well as FBank-83 (with CI) and FBank-80 (with CI). The architecture in all cases is L4 which achieved the best PER.

Fig. 5 shows the contribution of each phonetic class to the overall PER. As seen, applying monophones (CI) regularisation and adding pitch-related features [59] slightly reduces PER for most classes (except for nasals, semi-vowels and silence). It should be noted that the silence class inherently lacks pitch information, and regularisation with CI terms without proper tuning may result in over-regularisation.

We will use FBank83-L4-with-monophone hybrid system in the rest of this section as it achieves the best performance.

B. Recognition error per phonetic class

Fig. 6 shows the recognition errors (PER, Sub, Del and Ins) for the 8-class broad phonetic classes. The vowels contribute most to the overall PER, accounting for 6% of the total 14.1% PER on the test set. Fig. 7 (a) shows the *relative* contribution of each broad phonetic class to the overall PER: vowels 42%, plosives 15.6%, semi-vowels 12.5%, fricatives 11.2%, nasals 6.4%, diphthongs 5.9%, silence 4.6% and affricates 1.7%.

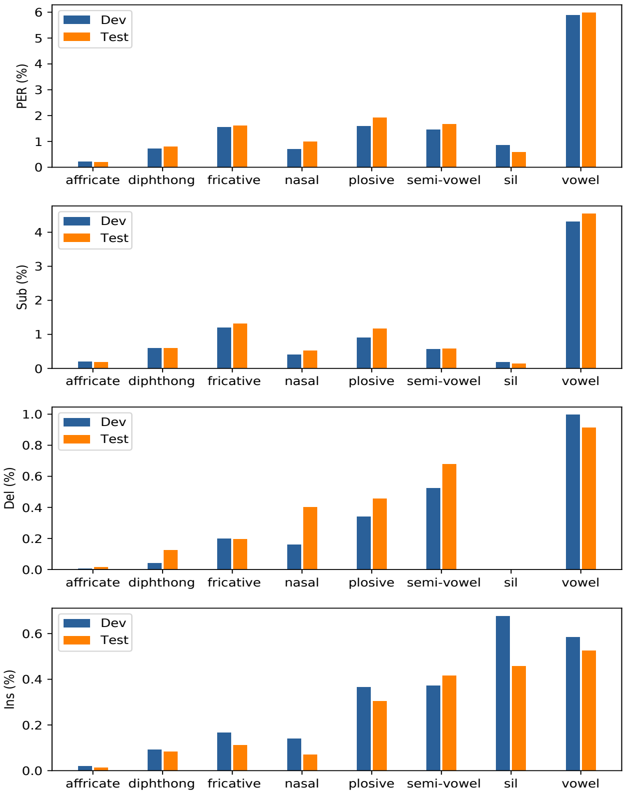


Fig. 6. Recognition errors on TIMIT's dev/test sets for the 8-class category.

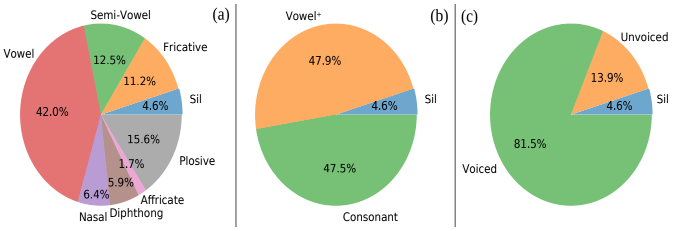


Fig. 7. Relative contribution of each broad class on TIMIT's test set PER.

Why do vowels account for the highest portion of PER? Interestingly, a similar observation was made in human phone recognition experiments in [69]. Vowel duration is significantly influenced by the speaker's speaking rate (tempo) [70]. Furthermore, as mentioned in Section I, vowels are the most useful phonetic units for speaker [5], [6] and emotion [13], [14] recognition tasks. These findings strongly suggest that vowels carry substantial speaker-related information, leading to heightened sensitivity to individual speakers and, consequently, contributing to greater PER.

Figs. 8 and 9 show PER along with the Sub/Del/Ins errors for the {consonant, vowel$^+$, silence} and {voiced, unvoiced, silence} categories. As seen, the contribution of the consonants and vowel$^+$s to the overall PER is very similar. In terms of the Sub/Del/Ins errors, however, the contribution of the vowel$^+$s and consonants is different. While the substitution errors for the consonants is smaller, the deletion and insertion rates are noticeably larger than those of the vowel$^+$s. Also note that the substitution error for silence is very small, the deletion error is zero and the insertion error is relatively large.

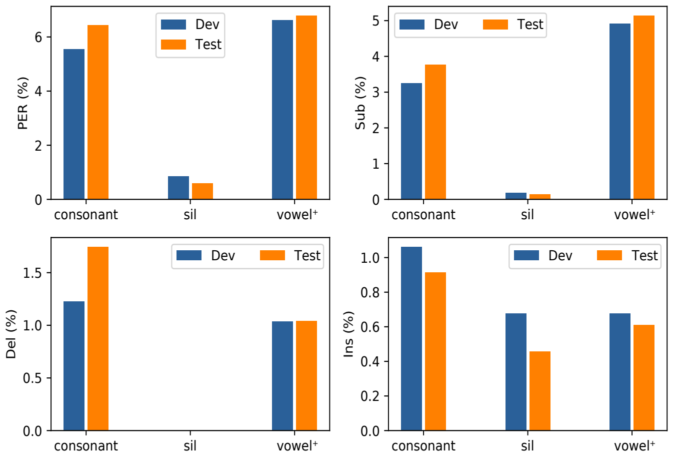


Fig. 8. PER, Sub, Del and Ins errors for the consonant and vowel$^+$ classes.

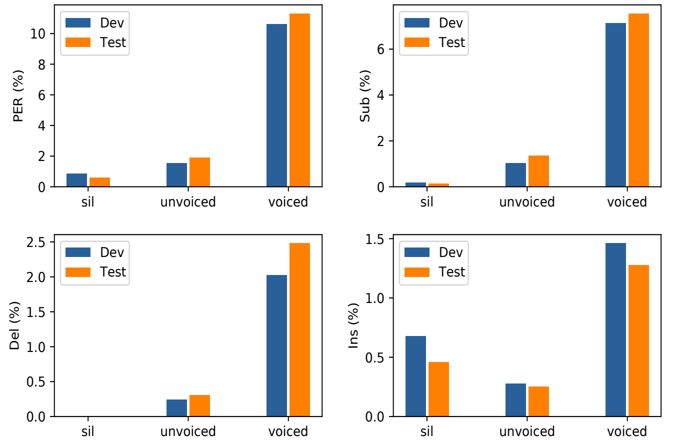


Fig. 9. PER, Sub, Del and Ins errors for the voiced and unvoiced classes.

As can be seen in Fig. 7 (b), the vowel$^+$s, consonants and silence classes account for 47.9%, 45.5% and 4.6% of the total PER, respectively. Similarly, the voiced class is responsible for 81.5% (=11.5/14.1*100) of the errors while the relative error due to unvoiced phones is 13.9% (Fig. 7(c)). Fig. 9 illustrates a remarkable contrast between the voiced and unvoiced categories with respect to Sub, Del, and Ins errors.

C. Dynamics of PER per phonetic class

Fig. 4 (a) depicts the dynamics of the overall PER vs epoch and Fig. 10 demonstrates the evolution of PER during training for the 8-class phonetic category. For a better visualisation, we dedicated an individual y-axis to the PER of each phonetic class. Similarly, Fig. 11 shows PER evolution for the voiced, unvoiced, consonant, vowel$^+$s and silence phonetic classes.

The temporal evolution of a performance metrics can be influenced by various factors such as the architecture, objective function, quality/amount of training data, and complexity of the classes being learned. As observed in Figs. 10 and 11, except for silence, the dynamics of the broad phonetic classes such as voiced and unvoiced or consonants and vowel$^+$s

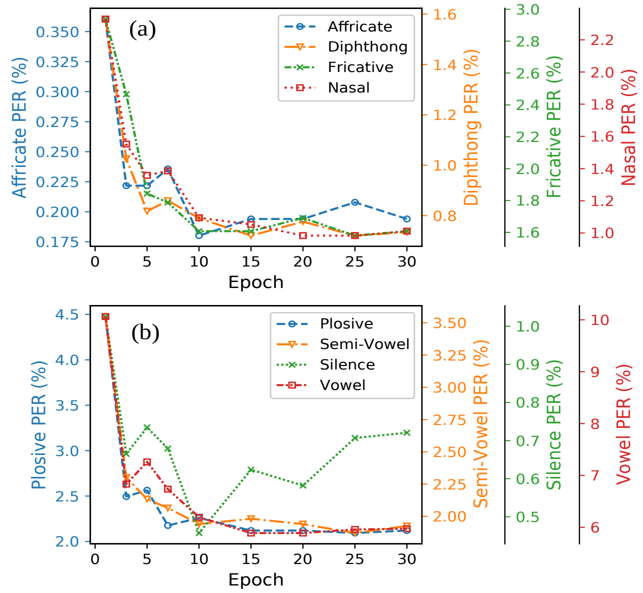


Fig. 10. Training dynamics (PER vs epoch) for the 8-class category.

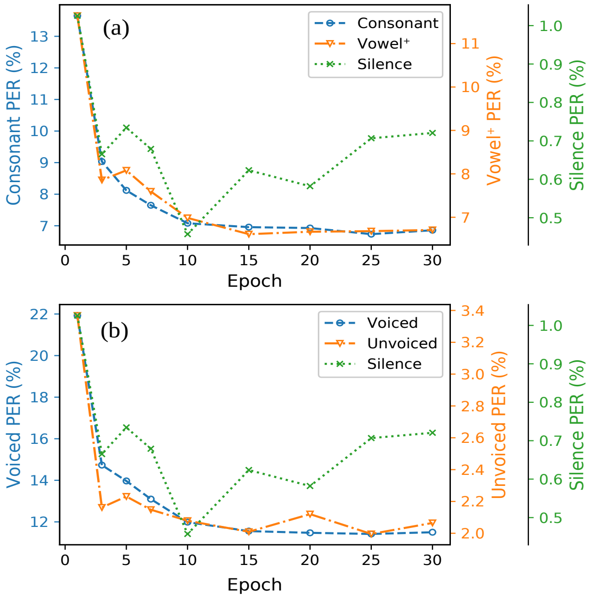


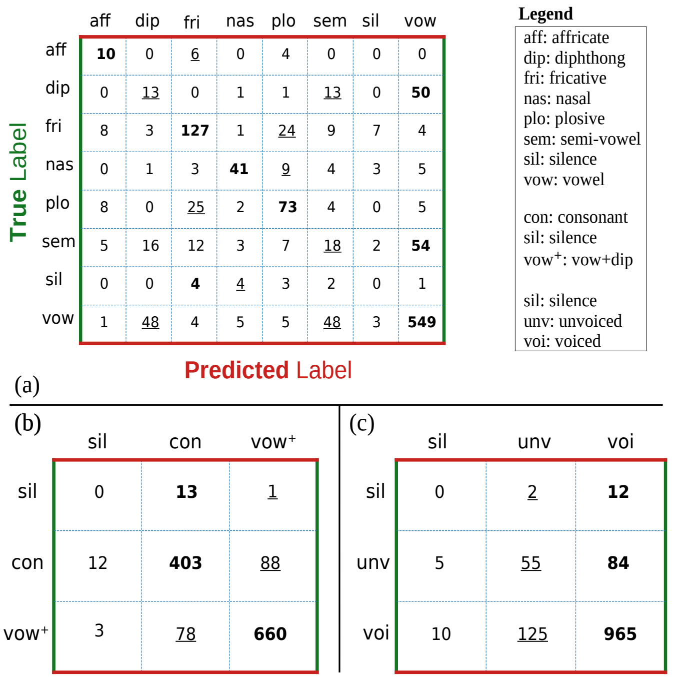
Fig. 11. Training dynamics (PER vs epoch) for the (a) {consonant, vowel$^+$, silence} and (b) {voiced, unvoiced, silence} broad phonetic categories.

are similar, despite differences between their acoustic characteristics, amounts of training data, the underlying learning complexities and PERs.

This interesting and rather counter-intuitive observation implies that the architecture and training objective – which are identical for all classes – play a major role in shaping the training dynamics, and the performance evolution is not significantly different across various classes.

D. Confusion Matrix

Now, we present and analyse the confusion matrices for the three broad phonetic classification. The confusion matrices are computed based on the substitution error which is the major component of PER (Figs. 4, 6, 8 and 9).


True Label
Predicted 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$^+$: vow+dip

sil: silence
unv: unvoiced
voi: voiced

(a)

(b)

	sil	con	vow$^+$
sil	0	**13**	1
con	12	**403**	88
vow$^+$	3	78	**660**

(c)

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

Fig. 12. Confusion matrices for TIMIT's dev set. The **bold** and underlined numbers denote the first and second mostly confused classes, respectively.

Fig. 12 shows the confusion matrices for TIMIT's dev[1] set over different broad phonetic definitions. The following observations can be made using Fig. 12 (a)

- affricates *are mostly confused with* (AMCW) themselves, fricatives and plosives, respectively;
- diphthongs AMCW vowels, themselves and semi-vowels;
- fricatives AMCW themselves and plosives;
- nasals AMCW themselves and plosives;
- plosives AMCW themselves and fricatives;
- semi-vowels AMCW vowels, themselves and diphthongs;
- silence is confused with fricatives, nasals and plosives;
- vowels AMCW themselves, diphthongs and semi-vowels.

Furthermore, based on Fig. 12 (b)

- silence is mostly confused with consonants and rarely confused with vowel$^+$s;
- consonants 80% (=403/(403+88+12)*100) of time are confused with themselves;
- vowel$^+$s 89% of time are confused with themselves.

Finally, as seen in Fig. 12 (c)

- silence is mostly confused with voiced phones;
- unvoiced phones are confused with the voiced and unvoiced phones 63% and 33% of time, respectively;
- voiced phones are confused with themselves and unvoiced phones 88% and 11.4% of time, respectively.

To analyse the confusions, we employ dimensionality reduction to a 2D space via LDA [33], using scikit-learn toolkit [71]. For each broad phonetic class, we plot a scatter graph in 2D along with an ellipse. To plot the ellipse, we first fit a 2D Gaussian with a full covariance matrix using the 2D features.

[1] Trend-wise similar observations were made for the test set. To save space, we only present the results for the dev set.

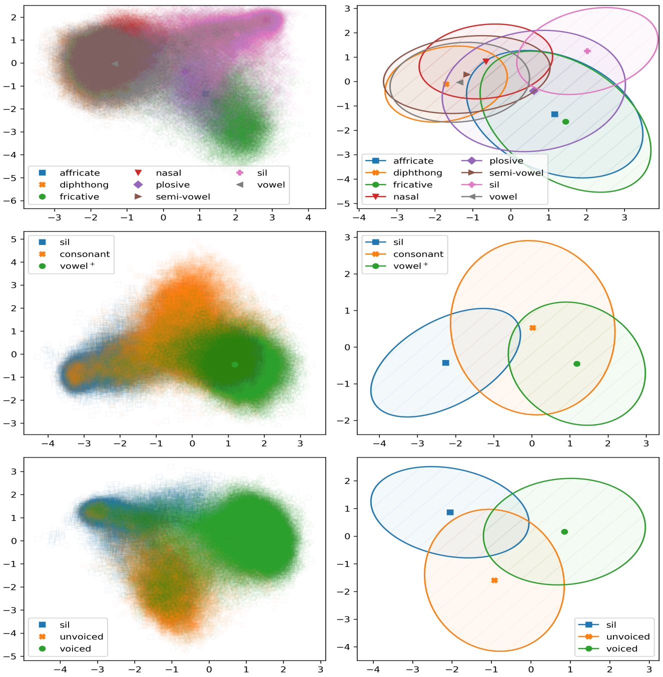


Fig. 13. Scatter plots at input level after mapping FBank-83 to 2D via LDA.

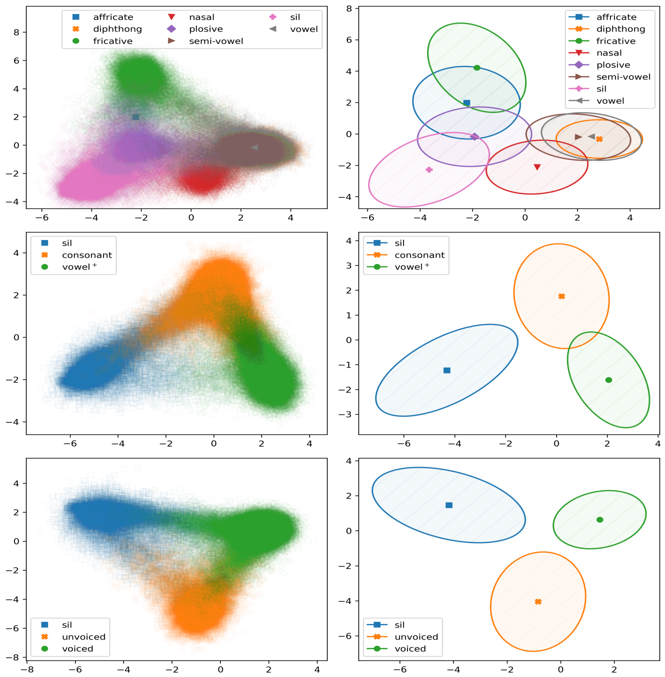


Fig. 14. Scatter plots at output level after mapping logits to 2D via LDA.

The ellipse's centre reflects the mean, its major and minor diagonals are aligned with the eigenvectors of the covariance matrix, whilst its semi-major and semi-minor axes are equal to the larger and smaller eigenvalues, respectively.

To gain insights into the confusions made by the acoustic model, we visualised the highest-level representations, namely the logits (where the final decisions are made by the softmax classifier) as well as the input, namely filterbank acoustic features, as shown in Figs. 13 and 14. Comparing these two figures not only helps explain the confusions, but also partially illustrates data processing along the acoustic model pipeline.

As observed in the scatter plots, the centroids of the phonetic classes at the output (logit) level are notably shifted away from each other, compared to the input (acoustic features) level. Additionally, at the logit level, each class occupies a more distinct and smaller neighborhood, indicating decreased eigenvalues. These enhance the linear separability and accuracy of the linear classification through softmax.

Another observation is that certain classes, such as vowels, diphthongs, and semi-vowels, remain located closely to each other at both low-level (acoustic feature) and high-level (logit) representations, with ellipses exhibiting similar orientations and closely located centers. Another example is affricates, which remain close to plosives and fricatives at both low and high levels. This suggests that although the pipeline partially disentangles the classes by pushing the centroids away from each other, it does not always result in an optimal separation, which would be desirable for achieving robust linear classification. Consequently, some low-level similarities persist even in the highest layers of the model, giving rise to undesired confusions between closely located classes.

It is worth noting that dimensionality reduction to 2D leads to some information loss which limits the ability to explain all observations. For instance, silence is mostly confused with the plosives and fricatives (Fig. 12 (a)). Although Fig. 14 depicts the proximity of the silence and plosives and explains this confusion, it fails in justifying the confusion between silence and fricatives as they appear to be far apart from each other.

The scatter plots in the left columns of Figs. 13 and 14 illustrate another interesting observation. For the {consonant, vowel$^+$, silence} and {voice, unvoiced, silence} broad phonetic classes, both at the input and output levels, the 2D data form clouds within a triangle. At the input level, the 2D features are concentrated in the center of the triangle, with heavy overlap among classes. However, at the output level, the classes are mostly located at the vertices of the triangle, with considerably lower overlap. Such a disentanglement greatly facilitates the linear classification.

Finally, as shown in Fig. 12 (b) and (c), the silence class tends to be confused with the consonants and voiced phonetic classes, which is challenging to explain. One would intuitively assume that silence should be more frequently confused with unvoiced consonants, as both silence and unvoiced sounds share similarities in their turbulent and noise-like excitation, in contrast to the quasi-periodic excitation of the voiced sounds.

We put forward two explanations for such a counter-intuitive observation. First, upon closely examining the decoding files, we found that silence is primarily confused with /n/ (nasal, voiced, consonant), /v/ (fricative, voiced, consonant) and /dx/ (plosive, voiced, consonant) phones and rarely confused with any other voiced consonant. These particular voiced consonants may exhibit features that are similar to silence, leading to misclassification. Second, as demonstrated in Fig. 2, approximately 64% of the training data belongs to the voiced class. This imbalance in data distribution could bias the model towards predicting more voiced sounds because regardless of

TABLE IV
PER FOR VARIOUS SYSTEMS. BASELINE IS FBANK83-L4-WITH-MONO. WSJ* DENOTES BASELINE SYSTEM PRE-TRAINED ON WSJ.

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

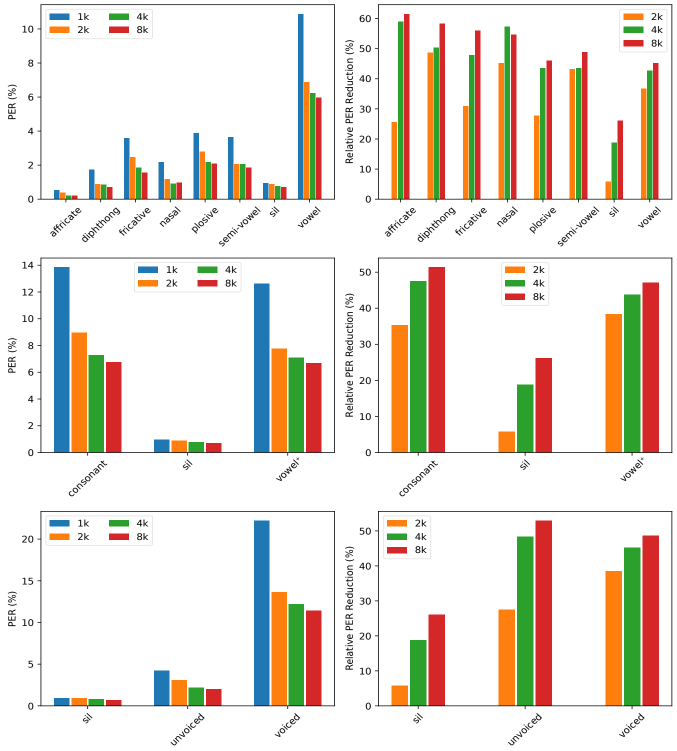


Fig. 15. (left column) PER for 1, 2, 4 and 8 kHz subbands. (right column) Relative gain with respect to PER of 1 kHz.

their accuracy, it results in a smaller overall training loss.

V. Comparing Different Systems

In this section, we compare the baseline hybrid FBank83-L4-with-mono system with other DNN and GMM based hybrid, as well as end-to-end (E2E) models. Table IV shows the PER of various systems discussed in this section. More details about each one is presented in the corresponding subsection.

A. Importance of different subbands

To study the importance of the 1, 2, 4 and 8 kHz subbands in the PER of BPCs, we trained phone recognisers using only the first 26, 40, 60 and 80 filters of the filterbank, respectively. All subband features were appended by the pitch-related features.

The left column of Fig. 15 shows PER for different phonetic classes and the right column illustrates the relative gain with respect to the 1 kHz system. Incorporating additional subbands provides extra information and contributes to improving PER for all phonetic classes. The only exception to this is the nasals, where the 4-kHz system outperforms the 8-kHz one.

There is a notable difference in the relative gain across various classes. The non-silence speech classes (voiced, unvoiced, consonant, and vowel^{+}) achieve a larger relative gain compared to silence, with a significant margin. The silence class is typically characterised by low energy in all frequency bands. Therefore, the importance of higher spectral subbands for this class is minimal. On the other hand, the relative gain in performance after inclusion of higher frequency subbands is larger for the unvoiced and consonant categories than for the voiced and vowel classes. This suggests that the high frequency spectral components are more discriminative and informative in recognising the unvoiced and consonant classes.

B. Unidirectional vs bidirectional LSTMs

Next, we look into the effectiveness of bidirectional vs unidirectional sequential modeling via LSTMs and how it affects the PER per phonetic class. Fig. 16 illustrates the PER for these systems, along with the relative gain of the bidirectional modelling with respect to the unidirectional one. As seen, bidirectional modeling provides significant improvement for all classes. The greatest improvement observed in diphthongs and fricatives, while the silence class benefits the least.

Quantitative comparison of the PERs of the silence with other classes (Fig. 16 (d) and Fig. 16 (f)) is insightful: while the relative gain for silence is around 4%, for others it is more than 20%. We hypothesise that the silence benefits the least from the sequential modelling due to its minimal susceptibility to contextual and neighboring phones (coarticulation).

C. Effect of noise (TIMIT vs NTIMIT)

Now, we examine the noise impact on PER of broad phonetic classes. To address the issue of variability in noise types and ensure reproducibility, instead of synthetically adding noise, we used the Network TIMIT (NTIMIT) corpus [35].

NTIMIT was collected by creating a *loopback*[2] telephone path to geographically distributed central offices in order to simulate different real-world local and long-distance telephone networks. This process introduces various noises including transmission and coding distortions. Although the severity of such noise is variable and often unpredictable due to differing line and telephone network conditions, NTIMIT is orthographically and phonetically equivalent to TIMIT.

For acoustic modelling, L4 architecture along with FBank-83 features and monophone regularisation were used. We achieved highly competitive PERs of 19.2% and 20.1% for the dev and test sets of NTIMIT, respectively.

Fig. 17 displays PER per phonetic class for TIMIT and NTIMIT along with the relative PER increase in NTIMIT, compared to TIMIT. It is evident that the performance drops for all phonetic classes with the largest relative PER increase observed for the fricatives (95%), nasals (72%) and plosives

[2] The transmitted audio signal was routed back to its original terminal.

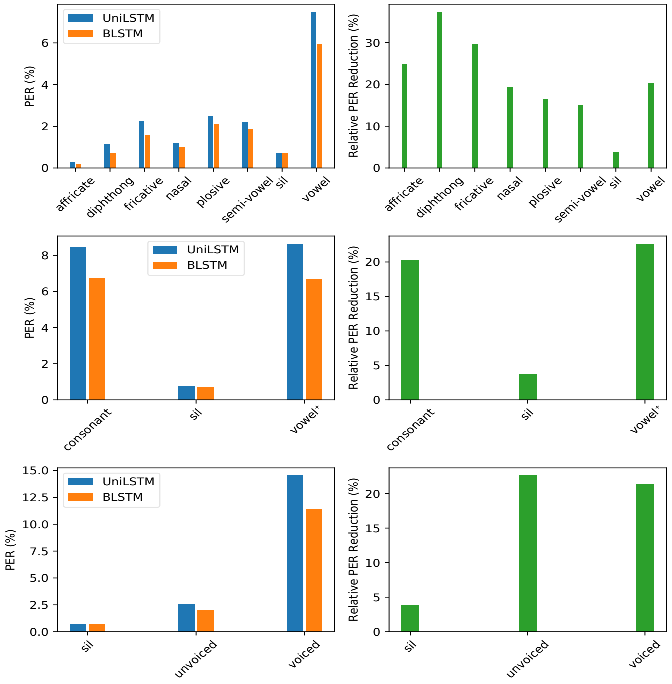


Fig. 16. (left column) Effect of applying unidirectional (UniLSTM) and bidirectional (BLSTM) LSTMs on PER per phonetic class. (right column) Relative PER reduction after replacing unidirectional with bidirectional LSTMs.

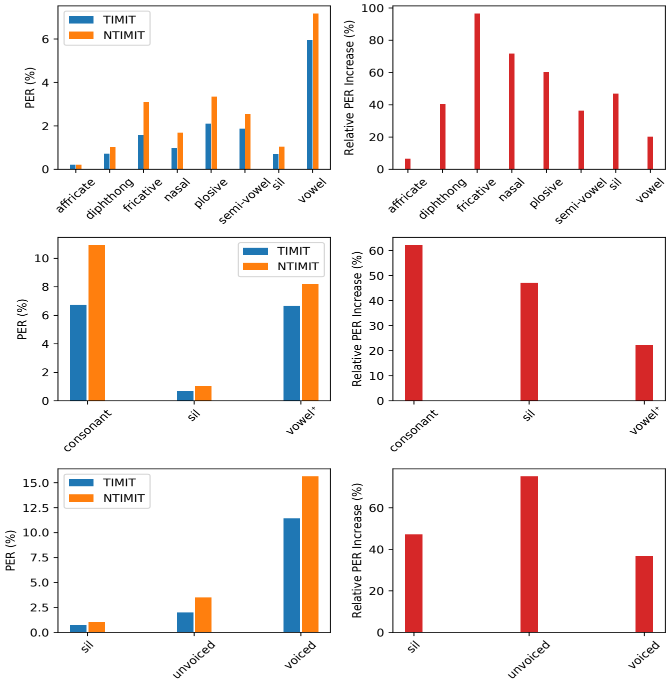
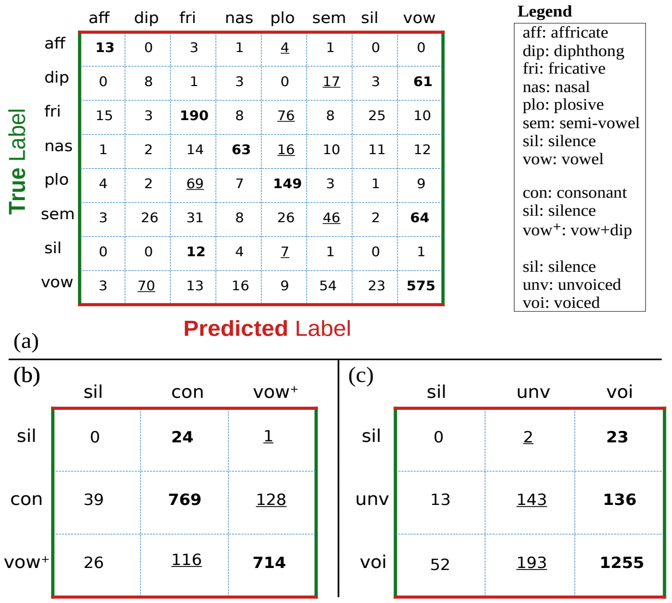


Fig. 17. TIMIT vs NTIMIT for different phonetic classes. The right column displays the relative PER increase on NTIMIT with respect to TIMIT.

(60%), respectively. The vowel+ and voiced phones appear to be more robust compared with the consonants and unvoiced classes. The relative PER elevation for the silence is about 48% while for the vowel+, consonant, voiced and unvoiced phones is 21%, 62%, 39% and 68%, respectively.

The vowel+ and voiced classes demonstrate greater robustness due to their larger energy, leading to a higher segmental signal-to-noise ratio at the corresponding frames. Moreover, when transmitting signals through a telephone network, the limited bandwidth predominantly impacts phones with higher frequency components like fricatives and plosives. As a result, the vowel+ and voiced classes are less affected since the majority of their spectral content falls within the telephone bandwidth, while consonants and unvoiced classes experience greater distortion owing to substantial spectral density located outside the bandwidth of the telephone network.


True Label
Predicted Label

(a)

	aff	dip	fri	nas	plo	sem	sil	vow
aff	**13**	0	3	1	4	1	0	0
dip	0	8	1	3	0	17	3	**61**
fri	15	3	**190**	8	76	8	25	10
nas	1	2	14	**63**	16	10	11	12
plo	4	2	69	7	**149**	3	1	9
sem	3	26	31	8	26	46	2	**64**
sil	0	0	**12**	4	7	1	0	1
vow	3	70	13	16	9	54	23	**575**

Legend
aff: affricate
dip: diphthong
fri: fricative
nas: nasal
plo: plosive
sem: semi-vowel
sil: silence
vow: vowel

con: consonant
sil: silence
vow+: vow+dip

sil: silence
unv: unvoiced
voi: voiced

(b)

	sil	con	vow+
sil	0	**24**	1
con	39	**769**	128
vow+	26	116	**714**

(c)

	sil	unv	voi
sil	0	2	**23**
unv	13	143	**136**
voi	52	193	**1255**

Fig. 18. Confusion matrices for NTIMIT's dev set. The **bold** and underlined numbers denote the first and second mostly confused classes, respectively.

Fig. 18 depicts the confusion matrices for NTIMIT. As seen, the overall confusion patterns (the bold and underlined items) remain similar to TIMIT (Fig. 12).

D. *GMM-HMM vs DNN-HMM*

In this subsection, we compare the baseline phone recogniser with a GMM-HMM system to analyse the errors and investigate which phonetic classes benefit further/less by replacing the GMMs with DNNs. The GMM-HMM system was built by Kaldi [53] using speaker adaptive training (SAT) [72], max-likelihood linear transformation (MLLT) [73], and LDA. Fig. 19 shows PERs along with the relative gains. As seen,

- ranking of PERs of different phonetic classes remains almost identical;
- the largest PERs correspond to vowels, plosives, fricatives and semi-vowels;
- vowel+s and consonants have similar PERs;
- PER of voiced class is notably larger than unvoiced one.

Comparing the relative PER reductions (right column in Fig. 19) shows that the maximum gain is achieved for the silence. While the relative gain varies in the range of 25% (semi-vowels) to 40% (plosives and fricatives), the relative PER reduction for the silence class reaches 55%. Compared to other phonetic classes, recognising the silence relies more heavily on the acoustic model (AM) as it is not effectively

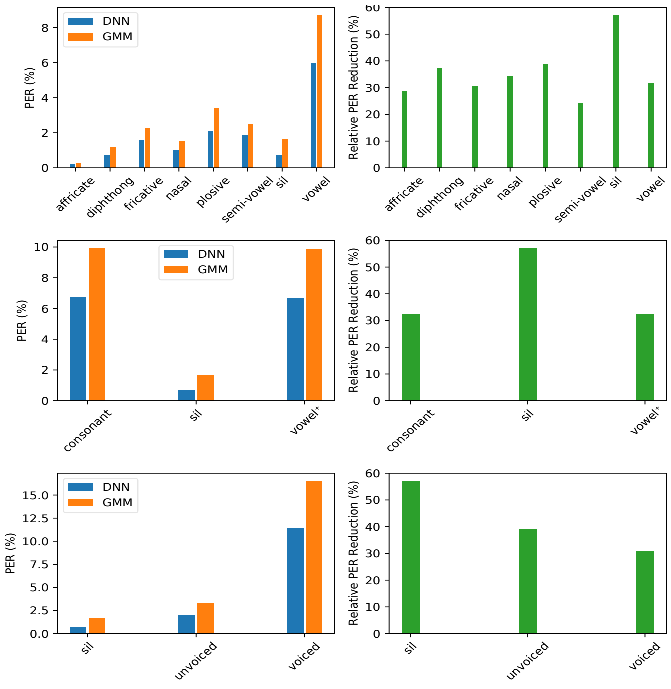
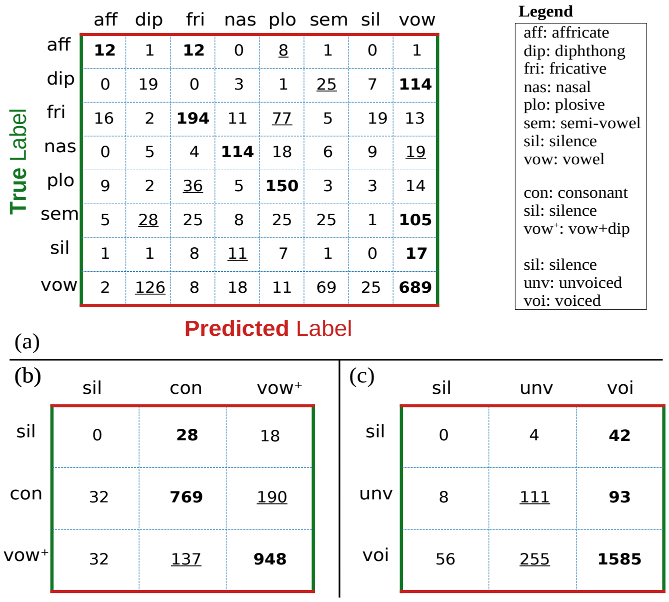


Fig. 19. DNN-HMM (L4) vs GMM-HMM (SAT-MLLT-LDA) hybrid systems for different phonetic classes. The right column shows the relative PER reduction on the DNN-HMM system with respect to the GMM-HMM one.


True Label
Predicted Label
(a)
(b)
(c)

(a)

	aff	dip	fri	nas	plo	sem	sil	vow
aff	**12**	1	**12**	0	<u>8</u>	1	0	1
dip	0	19	0	3	1	<u>25</u>	7	**114**
fri	16	2	**194**	11	<u>77</u>	5	19	13
nas	0	5	4	**114**	18	6	9	<u>19</u>
plo	9	2	<u>36</u>	5	**150**	3	3	14
sem	5	<u>28</u>	25	8	25	25	1	**105**
sil	1	1	8	<u>11</u>	7	1	0	**17**
vow	2	<u>126</u>	8	18	11	69	25	**689**

(b)

	sil	con	vow⁺
sil	0	**28**	18
con	32	**769**	<u>190</u>
vow⁺	32	<u>137</u>	**948**

(c)

	sil	unv	voi
sil	0	4	**42**
unv	8	<u>111</u>	**93**
voi	56	<u>255</u>	**1585**

Legend

aff: affricate
dip: diphthong
fri: fricative
nas: nasal
plo: plosive
sem: semi-vowel
sil: silence
vow: vowel

con: consonant
sil: silence
vow⁺: vow+dip

sil: silence
unv: unvoiced
voi: voiced

Fig. 20. Confusion matrices for the GMM-HMM system on TIMIT's dev set. The **bold** and <u>underlined</u> denote the first and second mostly confused classes.

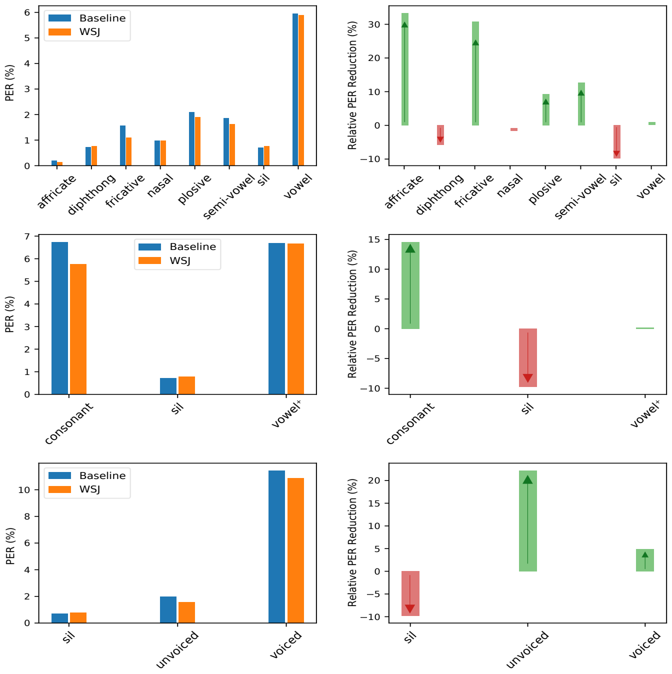


Fig. 21. Baseline vs the same system pre-trained on WSJ with frozen layers except for the output layer which is trained from scratch. The right column shows the relative PER reduction after transfer learning relative to baseline.

handled by the language model (LM). That is, the LM training data primarily consists of the text sources, least helpful in modelling the silence, especially the inter-word silence. Therefore, using a stronger AM is most beneficial for the silence class.

Fig. 20 shows the confusion matrix for the GMM-HMM system. Although many confusion patterns (the bold and underlined items) remain similar, there are some differences. For example, the silence class is mostly confused with fricatives and plosives in the DNN-based system (Fig. 12), while in the GMM-based one it is mostly confused with vowels and nasals. Also, vowels are equally confused with diphthongs and semi-vowels in the DNN-based system while in the GMM-based systems vowels are confused with diphthongs almost twice as many times as with semi-vowels.

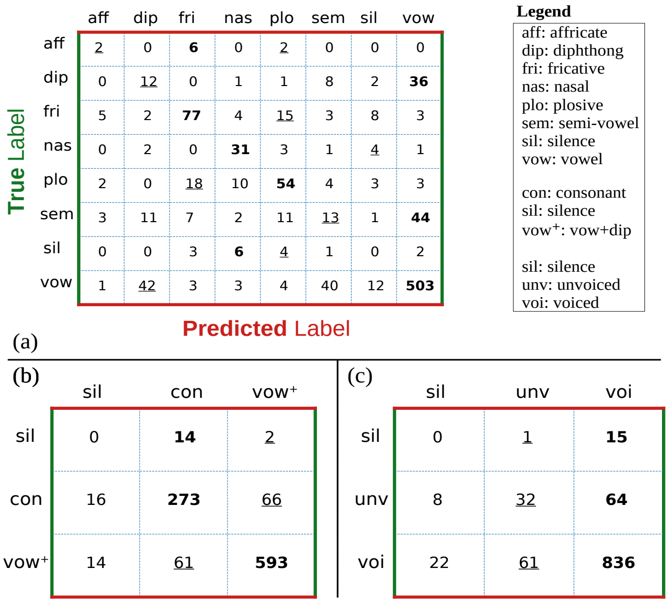
E. Transfer learning from WSJ

We also investigated the effect of transfer learning from WSJ. We first trained the baseline model on WSJ and then transferred all the weights, except for those between the penultimate and output layers which were trained on TIMIT from scratch. As observed in Table IV, the transfer learning from WSJ leads to relative performance gain of 10.1% and 7.1% on the dev and test sets, respectively.

Analysing PER for each BPC shows that although such a transfer learning is helpful for most classes, it does not improve PER for all classes. Specifically, the performance gets significantly worse for the silence class by -10% (relative).

Silence, due to its inherent variability, complexity, and acoustic properties that are strongly influenced by background noise and/or recording setup, can exhibit notable differences over various datasets. Consequently, the system trained on WSJ may not effectively generalise to the silence segments in TIMIT, resulting in increased errors in silence recognition.

On the other hand, while the performance remains almost unchanged for vowels with the relative PER reduction of 1%, the relative PER reduction for consonants is substantial and reaches 14.6%. We will discuss this in the next subsection.


True Label
Predicted Label
(a)
(b)
(c)

	aff	dip	fri	nas	plo	sem	sil	vow
aff	2	0	**6**	0	2	0	0	0
dip	0	12	0	1	1	8	2	**36**
fri	5	2	**77**	4	15	3	8	3
nas	0	2	0	**31**	3	1	4	1
plo	2	0	18	10	**54**	4	3	3
sem	3	11	7	2	11	13	1	**44**
sil	0	0	3	**6**	4	1	0	2
vow	1	42	3	3	4	40	12	**503**

Legend
aff: affricate
dip: diphthong
fri: fricative
nas: nasal
plo: plosive
sem: semi-vowel
sil: silence
vow: vowel

con: consonant
sil: silence
vow$^+$: vow+dip

sil: silence
unv: unvoiced
voi: voiced

(b)

	sil	con	vow$^+$
sil	0	**14**	2
con	16	**273**	66
vow$^+$	14	61	**593**

(c)

	sil	unv	voi
sil	0	1	**15**
unv	8	32	**64**
voi	22	61	**836**

Fig. 22. Confusion matrices on TIMIT's dev set after transfer learning from WSJ and training output layer from scratch. The **bold** and underlined numbers indicate the first and second mostly confused classes, respectively.

F. End-to-end vs hybrid

Now, we compare the baseline hybrid system with two end-to-end (E2E) models: Conformer [36] and wav2vec 2.0 (W2V) [37]. The Conformer and wav2vec 2.0 systems were trained by ESPnet [74] and FAIRSEQ [75], respectively. The decoding process did not involve any external language model.

The Conformer system was built from scratch using the TIMIT training data. The loss function includes two components: a cross entropy loss on top of the Conformer's decoder as well as a CTC loss on top of the Conformer's encoder. The former was used for decoding and the latter for regularisation purposes. Both components were weighted equally with a scaling factor of 0.5. To construct an effective model, we conducted experiments with different numbers of encoder and decoder layers. Among these, the model with eight conformer encoder layers and one conformer decoder layer demonstrated the best performance on the dev set. It achieved 18.2% and 20.0% PERs on the dev and test sets, respectively.

The W2V-large model was pre-trained on 60k hours of LibriVox data, in the self-supervised learning mode. Then, similar to [37]'s recipe, we frozen the convolutional layers, fine-tuned the transformer layers and trained (from scratch) a single feed-forward projection layer with 1024 nodes along with a CTC [76] loss on TIMIT phone recognition task, using Adam [77] optimiser. This system achieved state-of-the-art 7.1% and 8.3% PERs on the dev and test sets, respectively.

Fig. 23 illustrates the phone error rates per phonetic class. The Conformer-based E2E model shows inferior results compared to the baseline hybrid system, while the W2V system demonstrates remarkably higher performance across all broad phonetic categories. The only exception to this is the silence were the hybrid system outperforms the W2V based phone recogniser. This observation can be justified using the discussion put forward in Subsection V-E about the silence class.

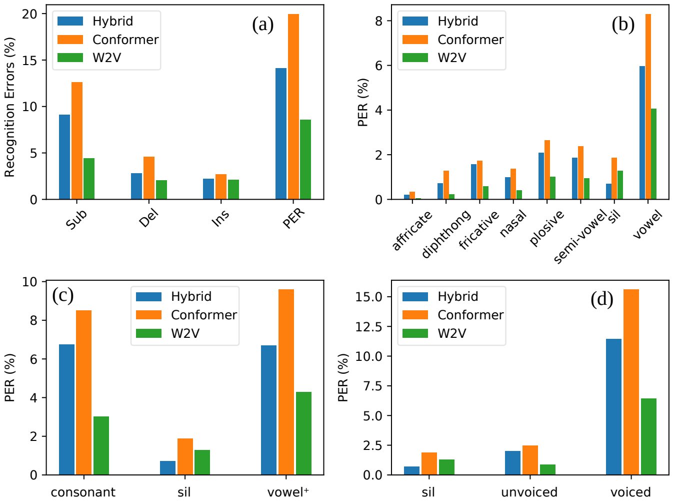


Fig. 23. Performance of the baseline hybrid system vs Conformer and wav2vec 2.0 (W2V) end-to-end phone recognisers over various BPCs.

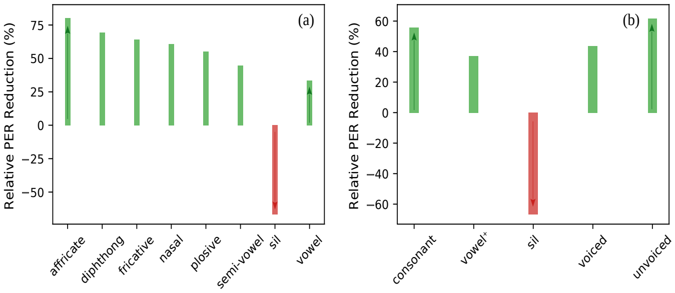


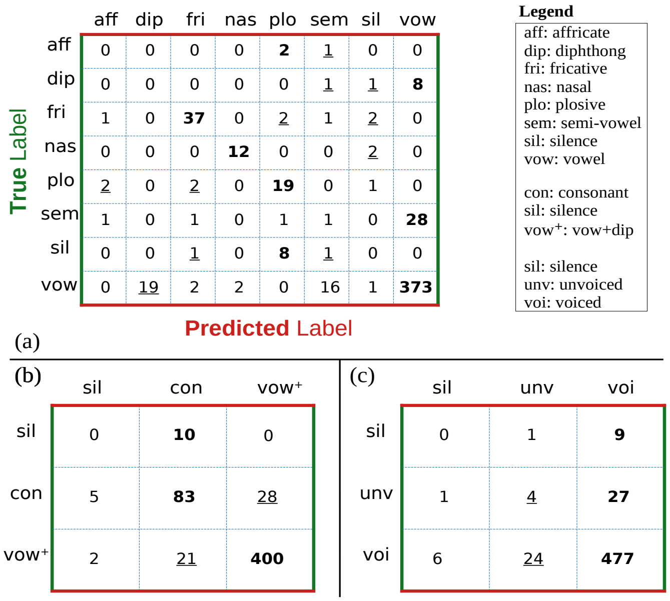
Fig. 24. Relative (to baseline) PER reduction after using W2V model.

Fig. 23 also shows that the hybrid and W2V systems render a similar performance in terms of the insertion error. From the deletion error viewpoint, W2V is slightly better. It, however, substantially outperforms the hybrid model in terms of the substitution error which is the major components of PER.

The rankings of the phonetic classes in terms of PER in the E2E systems remain similar to the hybrid one, e.g., the vowels, plosives, semi-vowels and fricatives still have the highest PERs. Also, the vowel$^+$ and voiced phones have a larger PER than the consonant and unvoiced classes, respectively.

Fig. 24 shows the relative PER reduction for various broad phonetic classes when using W2V model, relative to the baseline hybrid L4 system. Similar to the pre-training with WSJ, the maximum relative gain belongs to the affricate and unvoiced classes while the performance gets remarkably worse for the silence class (-67%). Additionally, the relative PER reduction for the vowel$^+$ (37%) is significantly less than that of the consonant (56%) class, consistent with the observation made after transfer learning from WSJ, where the performance gain for consonants was larger than vowels (Fig. 21).

This important observation suggests that increasing the amount of training data has a greater positive impact on consonants than on vowels. We propound two explanations for this: first, consonants often exhibit a greater variability owing to including a larger number of classes and having shorter du-


True Label
Predicted Label
(a)
(b)
(c)

	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**

Legend

aff: affricate
dip: diphthong
fri: fricative
nas: nasal
plo: plosive
sem: semi-vowel
sil: silence
vow: vowel

con: consonant
sil: silence
vow$^+$: vow+dip

sil: silence
unv: unvoiced
voi: voiced

(b)

	sil	con	vow$^+$
sil	0	**10**	0
con	5	**83**	28
vow$^+$	2	21	**400**

(c)

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

Fig. 25. Confusion matrices for the W2V system. The **bold** and underlined items denote the first and second mostly confused classes, respectively.

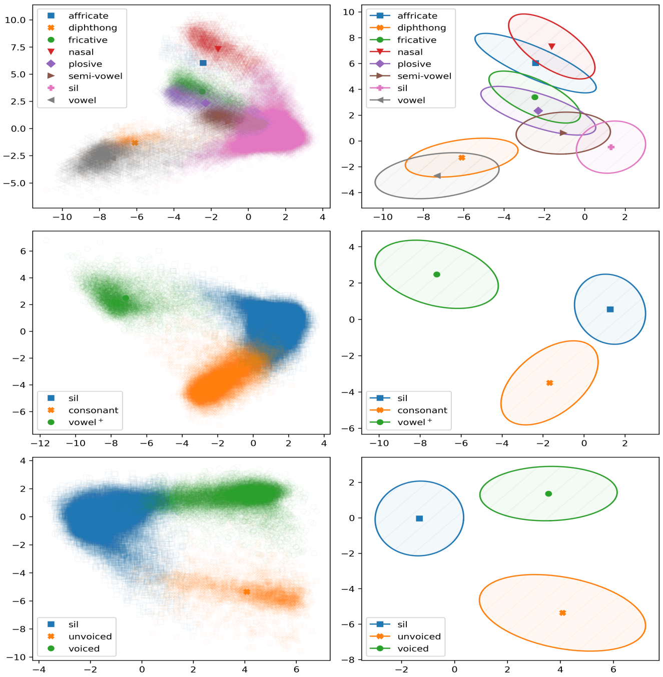


Fig. 26. Scatter plots at W2V's output level after mapping CTC logits to 2D via LDA. Compared to baseline model (Fig. 14), clusters are more distinct.

ration. Further, recognising consonants involves distinguishing between highly confusable voiced and unvoiced sounds (e.g., /z/ and /s/ or /b/ and /p/). These make learning consonants more data-intensive. Second, consonants are generally more frequent in natural languages than vowels, which implies that expanding the training data provides more exposure and learning opportunities for these phonetic classes.

We also studied the confusion matrices of the W2V system. Comparing Figs. 25 and 12 shows that in the W2V system the confusion matrix of the 8-class category has become sparser and, both intra- and inter-class confusions are dramatically reduced. In both systems, the intra-class confusion is dominant for the fricatives, nasals, plosives and vowels. Besides, semi-vowels and diphthongs are still mostly confused with vowels.

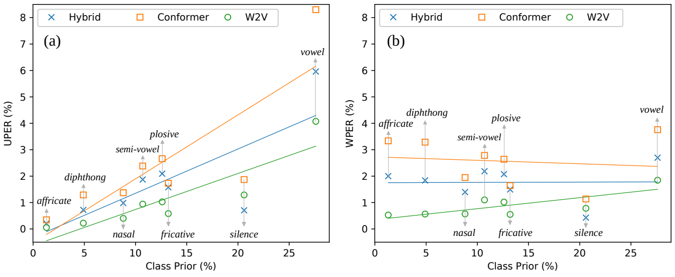


Fig. 27. Unweighted PER (UPER) and weighted PER (WPER) vs class priors for the baseline hybrid and end-to-end (Conformer and W2V) systems.

Fig. 26 presents the scatter plot of the CTC logits obtained from the fine-tuned W2V system after dimensionality reduction to 2D via LDA. A comparison with the logits of the baseline HMM-DNN system (Fig. 14) reveals more distinct clusters for different classes and reduced overlap. This improved class separation is highly desirable for classification purposes. Additionally, despite employing a deep structure like wav2vec-large, comprising of seven convolutional layers and 24 transformer encoder layers, acoustically similar classes such as plosives and fricatives or vowels and diphthongs remain closely grouped together, even at the logit level. This underscores the persistent challenge of disentangling acoustically similar phonetic units, even with advanced architectures.

G. PER vs class prior probabilities

Finally, we investigate the relationship between the amount of training data and PER across broad phonetic classes. Fig. 27 (a) shows PER vs class priors. To better understand and emphasise the trend, we conducted a linear regression.

Trend-wise, a higher class prior (indicating more training data) leads to a larger PER in all systems (hybrid and end-to-end). There are two exceptions to this, namely fricatives and silence but this trend is still counter-intuitive as one might expect more training data to result in a lower PER.

This is, however, the case in scenarios where the training data size/variability is expanded/enriched while the test set remains fixed, e.g., when applying data augmentation. For TIMIT, as demonstrated in Figs. 1 and 2, the phonetic class priors across the train, dev, and test sets are identical. This implies that when the amount of training data for a class is expanded, the amount of test data is increased proportionally as well. While more training data generally results in better learning, having more test data introduces larger variability. The observed trend in Fig. 27 (a) suggests the increase in the amount of training data alone is not sufficient to handle the complexity induced by the larger and possibly richer test set.

From the learning perspective, classes may vary in complexity, with certain classes being easier or more challenging to learn than others. This raises an important question: Do classes with higher error rates possess higher class complexity? We argue that assessing class complexity solely based on per-class error values, without considering class priors, is imprecise.

If the amount of training data per class were the same (uniform class priors) and the data quality and diversity were assumed to be consistent across all classes, then each class would have equal learning opportunities. In such a scenario, the per-class error can more accurately reflect the class complexity. However, in the current context, we lack a uniform prior as depicted in Figs. 1 and 2. Note that the non-uniform prior is not a deficiency for TIMIT, as it aims to sample the English language, where the frequency and importance of various phones are inherently different [55], [56], [57].

To account for the data imbalance, one solution is to normalise the PER per class with the corresponding prior probability. It involves weighting the recognition errors inversely proportional to the class priors. To this end, we define a weighted version of the PER,

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

$$WPER = \sum_{c=1}^{C} WPER_c = \sum_{c=1}^{C} \frac{Sub_c + Del_c + Ins_c}{N\ P_c\ C} \quad (2)$$

where P_c, Sub_c, Del_c and Ins_c indicate class prior, substitution, deletion and insertion for the phonetic class c; N is the total number of reference tokens, C is number of classes and, WPER and UPER denote the weighted and unweighted (typical) PERs, respectively. When the data is balanced, the prior probability mass function is uniform ($P_c = 1/C$) and, both weighted and unweighted PERs would be equal.

As seen in Fig. 27 (b), upon normalising errors with the class priors, WPER for various classes becomes comparable, and trend-wise varies in a very narrow range. However, even after considering the class prior, the vowel class still exhibits the largest PER, implying it holds the highest class complexity, owing to reasons discussed in Section IV-B.

VI. Conclusions and Scope for Future Work

In this paper, we investigated the performance of the TIMIT-based phone recognition systems beyond the commonly used phone error rate (PER) metric. PER shows the average performance and does not provide further details about the errors made by individual phonetic classes. We decomposed the overall substitution, deletion, insertion and PER and, computed each metric for three broad phonetic categories: {affricate, diphthong, fricative, nasal, plosive, semi-vowel, silence, vowel}, {consonant, vowel, silence} and {voiced, unvoiced, silence}. We investigated various hybrid (GMM-HMM and DNN-HMM) and end-to-end (Conformer and wav2vec 2.0) models and computed the performance metrics per phonetic class along with constructing a confusion matrix for each broad phonetic category. The confusion patterns were analysed using dimensionality reduction to 2D via LDA at both input (filterbank acoustic features) and output (logits) layers of the acoustic model. Moreover, we studied the effect of noise (NTIMIT), class priors, spectral subbands, bidirectional vs unidirectional sequential modelling and transfer learning (from WSJ). Finally, we compared and analysed the errors of the state-of-the-art hybrid and end-to-end systems.

The following summarises some of the key observations:

- training dynamics for all the phonetic classes is similar;
- replacing the uni- with bi-directional sequential modelling is least advantageous for silence;
- replacing GMMs with DNNs is most useful for silence;
- vowels and fricatives are the most and the least robust classes to noise (NTIMIT), respectively;
- pre-training (e.g., by WSJ or LibriVox) improves the performance over all phonetic classes except for silence;
- consonants benefit more than vowels from pre-training.

The proposed framework not only provides insights for conducting a more in-depth speech recognition error analysis and enhancing the interpretability and explainability of these DNN-based systems, but also has the potential to be leveraged in designing novel training regimes, loss functions, and performance evaluation metrics. Additionally, it contributes towards a better understanding of acoustic phonetics observations and phenomena. Future research directions could include exploring alternative modeling techniques, integrating linguistic context, and examining the impact of different languages, dialects, and speech styles on errors within the broad phonetic classes.

References

[1] K. Stevens, *Acoustic Phonetics*, ser. Current Studies in Linguistics Series. MIT Press, 2000.

[2] J. Yuan and M. Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *ICASSP*, 2010, pp. 4222–4225.

[3] B. Ludusan and E. Dupoux, "Automatic syllable segmentation using broad phonetic class information," *Procedia Computer Science*, vol. 81, pp. 101–106, 2016, sLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages.

[4] G. N. Clements, "The role of the sonority cycle in core syllabification," in *Papers in laboratory phonology I: Between the grammar and physics of speech*. Cambridge University Press, 1990.

[5] J. Eatock and J. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *ICASSP*, 1994, pp. 133–136.

[6] M. Antal and G. Toderean, "Speaker recognition and broad phonetic groups," in *SPPRA*, ser. SPPRA'06. ACTA Press, 2006, p. 155–159.

[7] Y.-J. Lu, C.-F. Liao, X. Lu, J. weih Hung, and Y. Tsao, "Incorporating Broad Phonetic Information for Speech Enhancement," in *INTERSPEECH*, 2020, pp. 2417–2421.

[8] G. Kubin and W. Kleijn, "Time-scale modification of speech based on a nonlinear oscillator model," in *ICASSP*, vol. i, 1994, pp. I/453–I/456 vol.1.

[9] O. Donnellan, E. Jung, and E. Coyle, "Speech-adaptive time-scale modification for computer assisted language-learning," in *Proceedings 3rd IEEE International Conference on Advanced Technologies*, 2003, pp. 165–169.

[10] G. Kubin, B. Atal, and W. Kleijn, "Performance of noise excitation for unvoiced speech," in *Proceedings., IEEE Workshop on Speech Coding for Telecommunications,*, 1993, pp. 35–36.

[11] L. Zhang, T. Wang, and V. Cuperman, "A CELP variable rate speech codec with low average rate," in *ICASSP*, vol. 2, 1997, pp. 735–738.

[12] T. Kempton and R. K. Moore, "Language identification: insights from the classification of hand annotated phone transcripts," in *Odyssey*. ISCA, 2008.

[13] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *INTERSPEECH*, 2004, pp. 889–892.

[14] F. Ringeval and M. Chetouani, "A vowel based approach for acted emotion recognition," in *INTERSPEECH*, 2008, pp. 2763–2766.

[15] J. Yuan, X. Cai, R. Zheng, L. Huang, and K. Church, "The role of phonetic units in speech emotion recognition," *ArXiv*, vol. abs/2108.01132, 2021.

[16] P. Scanlon, D. P. W. Ellis, and R. B. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 3, pp. 803–812, 2007.

[17] C. Lopes and F. Perdigão, "A hierarchical broad-class classification to enhance phoneme recognition," in *EUSIPCO*, 2009, pp. 1760–1764.

[18] ——, "Broad phonetic class definition driven by phone confusions," *Eurasip Journal on Advances in Signal Processing*, vol. 2012, no. 158, pp. —, July 2012.

[19] Y.-T. Lee, X.-B. Chen, H.-S. Lee, J.-S. R. Jang, and H.-M. Wang, "Multi-task learning for acoustic modeling using articulatory attributes," in *APSIPA*, 2019, pp. 855–861.

[20] T. N. Sainath and V. Zue, "A comparison of broad phonetic and acoustic units for noise robust segment-based phonetic recognition," in *INTERSPEECH*, 2008, pp. 2378–2381.

[21] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT '94. Association for Computational Linguistics, 1994, p. 307–312.

[22] T. N. Sainath, "Island-driven search using broad phonetic classes," in *ASRU*, 2009, pp. 287–292.

[23] G. Gravier and D. Moraru, "Towards phonetically-driven hidden markov models: Can we incorporate phonetic landmarks in hmm-based asr?" in *Non-Linear Speech Processing*, 2007.

[24] S. Ziegler, B. Ludusan, and G. Gravier, "Using broad phonetic classes to guide search in automatic speech recognition," in *INTERSPEECH*, 2012, pp. 1023–1026.

[25] L. Yang, J. Zhang, and Y. Yan, "Acoustic units selection in Chinese-English bilingual speech recognition," in *NOLISP*, 2007, pp. 96–99.

[26] A. Žgank, B. Horvat, and Z. Kačič, "Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity," *Speech Communication*, vol. 47, no. 3, pp. 379–393, 2005.

[27] R. Cole and V. Zue, "Speech as eyes see it," in *Attention and Performance VIII*, R. Nickerson, Ed. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1980, ch. 24.

[28] ——. Speech as eyes see it. Youtube. [Online]. Available: https://www.youtube.com/watch?v=cgUuUoqwGmA

[29] F. Pons and J. M. Toro, "Structural generalizations over consonants and vowels in 11-month-old infants," *Cognition*, vol. 116, no. 3, pp. 361–367, 2010.

[30] H. JR, B.-V. S, N. M, and M. J., "Consonants and vowels: different roles in early language acquisition," in *Developmental science*, vol. 14, no. 6, 2011, pp. 1445–58.

[31] T. Nazzi and A. Cutler, "How consonants and vowels shape spoken-language recognition," in *Annual Review of Linguistics*, vol. 5, no. 1, 2011, pp. 25–47.

[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993.

[33] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. [Online]. Available: probml.ai

[34] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *ICASSP*, 1992, pp. 899–902.

[35] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *ICASSP*, 1990.

[36] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020.

[37] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NIPS*, vol. 33, 2020, pp. 12 449–12 460.

[38] D. Byrd, "Relations of sex and dialect to reduction," *Speech Communication*, vol. 15, no. 1, pp. 39–54, 1994.

[39] A. K. Halberstadt and J. R. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *ICSLP*, 1998.

[40] Y. Hifny and S. Renals, "Speech recognition using augmented conditional random fields," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 17, pp. 354–365, 2009.

[41] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable CNNs with parametric modulated kernel-based filters," in *INTERSPEECH*, 2019.

[42] D. Oglic, Z. Cvetkovic, and P. Sollich, "Learning waveform-based acoustic models using deep variational convolutional neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 2850–2863, 2021.

[43] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005, iJCNN 2005.

[44] C. Glackin, J. Wall, G. Chollet, N. Dugan, and N. Cannings, "TIMIT and NTIMIT phone recognition using convolutional neural networks," in *Pattern Recognition Applications and Methods*, M. De Marsico, G. S. di Baja, and A. Fred, Eds. Springer International Publishing, 2019, pp. 89–100.

[45] J.-P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech communication*, vol. 51 4, pp. 352–368, 2009.

[46] A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *ICASSP*, 2014, pp. 5552–5556.

[47] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," *SLT*, pp. 1021–1028, 2018.

[48] M. T. S. Al-Kaltakchi, R. R. O. Al-Nima, M. A. M. Abdullah, and H. N. Abdullah, "Thorough evaluation of TIMIT database speaker identification performance under noise with and without the g.712 type handset," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 851–863, 2019.

[49] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 20, no. 1, pp. 14–22, 2012.

[50] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013, pp. 6645–6649.

[51] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[52] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[53] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[54] X. Huang, A. Acero, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[55] K. N. Stevens, *The quantal nature of speech: Evidence from articulatory-acoustic data*. University Park Press, 1972, pp. 384–425.

[56] J. S. Perkell, "Movement goals and feedback and feedforward control mechanisms in speech production," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 382–407, 2012.

[57] B. Lindblom, *Phonetic universals in vowel systems*. Academic Press, 1986, pp. 13–44.

[58] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *ICANN*, 2005.

[59] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," *ICASSP*, pp. 2494–2498, 2014.

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[61] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines." in *ICML*, 2010, pp. 807–814.

[62] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[63] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *ICASSP*, 2019.

[64] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Workshop on Autodiff*, 2017.

[65] T. Tieleman and G. Hinton, "Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning," *COURSERA Neural Networks Mach. Learn*, 2012.

[66] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," in *ICASSP*, 2015, pp. 4290–4294.

[67] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 876–890, 2023.

[68] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus," in *MLMI*, 2005.

[69] B. T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition depending on speech-intrinsic variability," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3126–3141, November 2010.

[70] M. Fourakis, "Tempo, stress, and vowel reduction in american english," *J. Acoust. Soc. Am.*, vol. 90, pp. 1816–1827, 1991.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[72] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," in *ICASSP*, vol. 2, 1997, pp. 1043–1046 vol.2.

[73] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

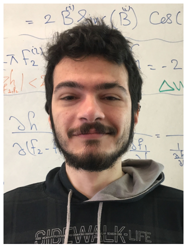
[74] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," *Proc. INTERSPEECH 2018*, p. 2207–2211, 2018.

[75] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[76] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*. ACM, 2006, pp. 369–376.

[77] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

Erfan Loweimi (S'10 — M'18) is a research associate with the Speech Group at the University of Cambridge and a visiting researcher at King's College London (KCL). He was a post-doctoral researcher at KCL (2021-2023) and at the Centre for Speech Technology Research (CSTR), University of Edinburgh (2018-2021). He received his B.Sc. (2007) in Electronics Engineering from Shahid Chamran University of Ahvaz, M.Sc. (2011) in Electronics Engineering from Amirkabir University of Technology (Tehran Polytechnic), and Ph.D. (2018) in Computer Science from the University of Sheffield. His research interests include acoustic modeling from raw signal representations, end-to-end ASR, robust model-based ASR, and phase-based speech processing. He is an associate member of IEEE Speech and Language Technical Committee (SLTC).



Andrea Carmantini (S'19) Andrea Carmantini is a PhD student at the Centre for Speech Technology Research of the University of Edinburgh. He received the B.A. in Asian Studies at Sapienza - University of Rome in 2015 and the M.Sc. in Speech and Language Processing at the University of Edinburgh. His research interests include methods for low resource speech recognition, model adaptation and semi-supervised training.

Peter Bell received the B.A. degree in mathematics in 2002 and the M.Phil. degree in computer speech, text and Internet technology in 2005 from the University of Cambridge, and the Ph.D. degree in automatic speech recognition from the University of Edinburgh, in 2010. He is a professor in speech technology with the School of Informatics, University of Edinburgh. His research interests include domain adaptation, regularisation, and low-resource methods for acoustic modeling.

Steve Renals (M'91 — SM'11 – F'14) received the B.Sc. degree in chemistry from the University of Sheffield, in 1986 and the M.Sc. degree in artificial intelligence in 1987 and the Ph.D. degree in neural networks and speech recognition from the University of Edinburgh, in 1991. He is Professor of speech technology with the School of Informatics, University of Edinburgh, having previously held positions at ICSI Berkeley, the University of Cambridge, and the University of Sheffield. His research interests include ASR, spoken language processing, and machine learning. Dr Renals is a fellow of ISCA (2016) and a Senior Area Editor of the IEEE OPEN JOURNAL OF SIGNAL PROCESSING.

Zoran Cvetkovic (Senior Member, IEEE) received the Dipl.Ing. and Mag. degrees from the University of Belgrade, the M.Phil. degree from Columbia University, and the Ph.D. degree in electrical engineering from the University of California, Berkeley. He is currently a Professor of Signal Processing with King's College London. He held research positions with EPFL (1996), and with Harvard University (2002–2004). Between 1997 and 2002, he was a member of the technical staff of AT&T Shannon Laboratory. His research interests are in the broad area of signal processing, ranging from theoretical aspects of signal analysis to applications in audio and speech technology. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.