

Chapter 2

Experimental Background

This chapter contains background information for the experimental work presented in this thesis. This includes information about the TIMIT and JUPITER databases, the SUMMIT speech recognition system, and the training of acoustic models.

2.1 The TIMIT Database

The TIMIT acoustic-phonetic continuous speech corpus [40] was recorded at Texas Instruments (TI), transcribed at the Massachusetts Institute of Technology (MIT), and verified and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST). It contains speech from 630 speakers representing 8 major dialect divisions of American English, each speaking 10 phonetically-rich sentences. There are 438 male speakers and 192 female speakers. The corpus includes the speech waveform files with corresponding time-aligned orthographic and phonetic transcriptions.

2.1.1 TIMIT Phones and Phone Classes

Table 2.1 shows the IPA and ARPAbet symbols for the 61 phones in the TIMIT corpus. The ARPAbet symbols will be used throughout the thesis. In accordance with common practice [43], we collapsed the 61 TIMIT labels into 39 labels before

IPA	ARPAbet	Example	IPA	ARPAbet	Example
[ɑ]	aa	<i>bob</i>	[ɪ]	ix	<i>debit</i>
[æ]	ae	<i>bat</i>	[i ^y]	iy	<i>beet</i>
[ʌ]	ah	<i>but</i>	[j̃]	jh	<i>joke</i>
[ɔ]	ao	<i>bought</i>	[k]	k	<i>key</i>
[ɑ ^w]	aw	<i>bout</i>	[k ^ɹ]	kcl	k closure
[ə]	ax	<i>about</i>	[l]	l	<i>lay</i>
[ə ^h]	ax-h	<i>potato</i>	[m]	m	<i>mom</i>
[ə ^r]	axr	<i>butter</i>	[n]	n	<i>noon</i>
[ɑ ^y]	ay	<i>bite</i>	[ŋ]	ng	<i>sing</i>
[b]	b	<i>bee</i>	[ɹ̃]	nx	<i>winner</i>
[b ^ɹ]	bcl	b closure	[o ^w]	ow	<i>boat</i>
[ç]	ch	<i>choke</i>	[o ^y]	oy	<i>boy</i>
[d]	d	<i>day</i>	[p]	p	<i>pea</i>
[d ^ɹ]	dcl	d closure	[pau]	pau	<i>pause</i>
[ð]	dh	<i>then</i>	[p ^ɹ]	pcl	p closure
[r]	dx	<i>muddy</i>	[ʔ]	q	glottal stop
[ɛ]	eh	<i>bet</i>	[r]	r	<i>ray</i>
[l]	el	<i>bottle</i>	[s]	s	<i>sea</i>
[m]	em	<i>bottom</i>	[ʃ]	sh	<i>she</i>
[n]	en	<i>button</i>	[t]	t	<i>tea</i>
[ŋ]	eng	<i>Washington</i>	[t ^ɹ]	tcl	t closure
[∅]	epi	epenthetic silence	[θ]	th	<i>thin</i>
[ɜ]	er	<i>bird</i>	[ʊ]	uh	<i>book</i>
[e ^y]	ey	<i>bait</i>	[u ^w]	uw	<i>boot</i>
[f]	f	<i>fin</i>	[ü]	ux	<i>toot</i>
[g]	g	<i>gay</i>	[v]	v	<i>van</i>
[g ^ɹ]	gcl	g closure	[w]	w	<i>way</i>
[h]	hh	<i>hay</i>	[y]	y	<i>yacht</i>
[ɦ]	hv	<i>ahead</i>	[z]	z	<i>zone</i>
[ɪ]	ih	<i>bit</i>	[ž]	zh	<i>azure</i>
-	h#	utterance initial and final silence			

Table 2.1: IPA and ARPAbet symbols for phones in the TIMIT corpus with example occurrences

scoring. The mapping is shown in Table 2.2. In addition, glottal stops were ignored for classification experiments, but included for recognition experiments. We have decided to divide the TIMIT phonetic labels into 6 broad, manner classes: vowels and semivowels (VS), nasals and flaps (NF), strong fricatives (SF), weak fricatives and aspirants (WF), stops (ST), and closures. Alternatively, we have partitioned the phones into three broad classes: sonorants (SON), obstruents (OBS), and silences (SIL). Table 2.3 shows the membership of each of these phonetic classes.

1	iy	20	n en nx
2	ih ix	21	ng eng
3	eh	22	v
4	ae	23	f
5	ax ah ax-h	24	dh
6	uw ux	25	th
7	uh	26	z
8	ao aa	27	s
9	ey	28	zh sh
10	ay	29	jh
11	oy	30	ch
12	aw	31	b
13	ow	32	p
14	er axr	33	d
15	l el	34	dx
16	r	35	t
17	w	36	g
18	y	37	k
19	m em	38	hh hv
39	bcl pcl dcl tcl gcl kcl q epi pau h# not		

Table 2.2: Mapping from 61 classes to 39 classes for scoring of results, after [43].

2.1.2 TIMIT Data Sets

The sentences in the TIMIT corpus are divided into three types: dialect (SA), phonetically-compact (SX), and phonetically-diverse (SI). The dialect sentences were designed to reveal the dialectical variation of the speakers, and were read by all 630

Phone Class	# TIMIT labels	TIMIT labels
Vowel/Semivowel (VS)	25	aa ae ah ao aw ax axh axr ay eh er ey ih ix iy ow oy uh uw ux el l r w y
Nasal/Flap (NF)	8	em en eng m n ng nx dx
Strong Fricative (SF)	6	s z sh zh ch jh
Weak Fricative (WF)	6	v f dh th hh hv
Stop (ST)	6	b d g p t k
Closure (CL)	9	bcl dcl gcl pcl tcl kcl epi pau h#
Sonorant (SON)	33	Vowel/Semivowel + Nasal/Flap
Obstruent (OBS)	18	Strong Fric + Weak Fric + Stop
Silence (SIL)	9	Same as Closure

Table 2.3: Phonetic subsets which will be used in subsequent experiments.

speakers. The two dialect sentences were “She had your dark suit in greasy wash water all year.” and “Don’t ask me to carry an oily rag like that.” The phonetically-compact (SX) sentences were hand-designed to be phonetically comprehensive as well as compact, in the sense of brevity. The phonetically diverse (SI) sentences were selected from existing text sources. Table 2.4 indicates the number of unique sentence orthographies of each type, the number of speakers per unique sentence orthography, and the number of sentences of each type spoken by each speaker.

Sentence Type	# Sentences	# Speakers/ Sentence	Total	# Sentences/ Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342	-	6300	10

Table 2.4: TIMIT speech material according to sentence type.

The core test set was selected to include 2 males and 1 female from each of the 8 dialect regions. Table 2.5 shows the 24 speakers in the core test set, along with their dialect region. There are 8 texts for each speaker (dialect sentences were excluded), for a total of 192 utterances in the core test set.

Dialect	Speakers
New England	mdab0 mwbt0 felc0
Northern	mtas1 mwew0 fpas0
North Midland	mjmp0 mlnt0 fpkt0
South Midland	ml10 mtl0 flm0
Southern	mbpm0 mkl0 fnlp0
New York City	mcmj0 mjdh0 fmgd0
Western	mgrt0 mnjm0 fdhc0
Army Brat (moved around)	mjl0 mpam0 fmld0

Table 2.5: 24 speakers in the TIMIT core test set, with their dialect region.

1	faks0	11	fdac1	21	fjem0	31	mgwt0	41	mjar0
2	mmdb1	12	mmdm2	22	mpdf0	32	fcmh0	42	fkms0
3	mbdg0	13	mbwm0	23	mcs0	33	fadg0	43	fdms0
4	fedw0	14	mgjf0	24	mglb0	34	mrtk0	44	mtaa0
5	mtdt0	15	mthc0	25	mwjg0	35	fnmr0	45	frew0
6	fsem0	16	mbns0	26	mmjr0	36	mdls0	46	mdlf0
7	mdvc0	17	mers0	27	fmah0	37	fdrw0	47	mrcs0
8	mrjm4	18	fcal1	28	mmwh0	38	fjsj0	48	majc0
9	mjsw0	19	mreb0	29	fgjd0	39	fjmg0	49	mroa0
10	mteb0	20	mjfc0	30	mrjr0	40	fmml0	50	mrws1

Table 2.6: 50 speakers in the TIMIT development set.

The NIST “complete” test set was formed by including all 7 repetitions of the SX texts in the core test set. This procedure resulted in adding another 144 speakers to the core set, for a total of 168 speakers in the complete test set. This set was not used in this thesis. The reason for this is that we made use of a development set which overlaps with this definition of the “complete” test set.

The NIST training set consists of the 462 speakers which are not included in either the “core” or “complete” test sets. With the exception of the dialect (SA) sentences, which are excluded from classification and recognition experiments, there is no overlap between the texts read by the training and testing speakers.

We made extensive use of a 50-speaker development set. The core set was reserved

Set	# Speakers	# Utterances	# Hours
Train	462	3,696	3.14
Development	50	400	0.34
Core Test	24	192	0.16
“Full” Test	118	944	0.81

Table 2.7: Number of speakers, utterances, and hours of speech in the TIMIT training, development, core test, and “full” test sets.

Phones	462-speaker Train	50-speaker Development	24-speaker Core	118-speaker “full” Test
VS	58,840	6,522	3,096	15,387
NF	14,176	1,502	731	3,566
SF	13,157	1,326	661	3,169
WF	8,990	1,014	467	2,323
ST	16,134	1,685	799	4,022
CL	28,928	3,008	1,461	7,230
glottal(q)	2,685	277	118	650
Total	142,910	15,334	7,333	36,347

Table 2.8: Token counts in phonetic subclasses for the TIMIT training, development, core test, and “full” test sets.

for final testing in order to avoid biasing results toward the core set. Thus, experiments for system design and modification were performed using the development set. The speakers in this set are disjoint from both the training set and the core test set. Table 2.6 lists the 50 speakers in the development set. In Chapter 3, we make use of a 118-speaker test set which consists of the “complete” test set, minus our development set. We refer to this 118-speaker set as the “full” test set.

Table 2.7 summarizes the number of speakers, the number of utterances, and the number of hours of speech in each of the sets used in the experiments in this thesis. Table 2.8 indicates the number of tokens in each of the data sets. These totals are helpful for the detailed results in Section 5.1.3, where the exact number of errors is reported along with the percent error.