Appendix A

SUN Speech database

A.1   Description

The SUN Speech database [12] was compiled by the Department of Electrical and Electronic Engineering of the University of Stellenbosch containing phonetically labelled speech in both Afrikaans and English. Speech data was recorded under controlled circumstances with 12 bit resolution and a 16kHz sampling rate. Details of the number of speakers and the number of sentences spoken by each group of speakers are given in Table A.1. The 60 sentences comprising the four sentence sets were chosen to exhibit the diversity of phonemes in the two languages.

Table A.1: Description of SUN Speech database: number of male and female speakers and total number of speakers for each sentence set

Language	Sentence set	Number of speakers			Number of sentences
		Male	Female	Total	
Afrikaans	1	24	16	40	10
	2	18	12	30	10
English	3	33	17	50	20
	4	22	4	26	20

A.2   Subdivision into training and test sets

The database is not entirely consistent in that some speakers, but not all, spoke sentences from more than one sentence set. For experimental purposes it is desired to obtain as much invariance as possible between the training set and the test set. A speaker independent, sentence independent division of the Afrikaans data can be obtained by using data from the first sentence set for training or adaptation and data from the second sentence set for testing. If data from the same speaker is available for both sentence sets, then data from only one of the sentence sets are used for either training or testing. A subset of the Afrikaans training set is also defined, containing the first 10 utterances from a group of 8 speakers who spoke all 20 Afrikaans sentences, with the second 10 sentences of these speakers making up the speaker dependent test set. The speaker dependent test set is not used in experiments reported in this thesis, except to create a pronunciation dictionary, but has been used previously by the authors for cross-language speaker and multispeaker adaptation experiments [33]. Details of the composition of the various subdivisions of the database given in Table A.2.

Table A.2: Subdivision of SUN Speech database into an Afrikaans training set, training subset, speaker dependent test set and speaker independent test set, as well as an English set

Language	Set	Speakers			Sentence	Label	Duration
		Male	Female	Total	numbers	count	(seconds)
Afrikaans	train	23	16	39	1-10	17251	1555
	train subset	2	6	8	1-10	3466	316
	SD test	2	6	8	11-20	5128	441
	SI test	14	1	15	11-20	9413	745
English		55	21	76	21-60	93778	7757

A.3   Phonetic content and labelling

A total of 59 phonetic categories, including both a *silence* and *unknown* category, were used to segment both the Afrikaans and the English speech. It was attempted to assign the labels phonetically, i.e. according to the sound produced, rather than phonologically assigning the labels, i.e. according to what was supposed to be said. The complete list of symbols along with the database numerical representation, computer phonetic representation, examples and frequency of occurrence is given in Tables A.3 and A.4. The frequency of occurrence is useful to evaluate the phonetic composition and diversity of the database, as well as the match between the two languages and also the match between the Afrikaans training and test sets.

It is evident from Tables A.3 and A.4 that many of the phonemes with no given English examples occur relatively frequently as labels in the English speech. Even though these phonemes do not possess accurately representative examples in English, they may occur due to particular pronunciations of certain words. This happens especially when rounding of the front vowels in English occurs, such as when [e] is rounded to form [ø] and [i] is rounded to form [y], which are not usually associated with English speech. The significant exceptions (more than 0.1% of all occurrences) and the words most commonly containing these labels are listed next, in order of decreasing occurrence.

y : educational, to, reputation, dreary

   : dead, dreary, various, yesterday

ø : guilty, beautiful, annual, continues

œ : to, of, will, a

œy : educational, motivate, reputation, observation

As far as the Afrikaans speech is concerned, there are also a few of the phonetic labels

Table A.3: Phonetic classes, labels, SUN Speech numbering and computer phonetic labels with English and Afrikaans examples, as well as the percentage occurrence relative to all labels of the label in the English data (F_E) and the Afrikaans training set(F_{Atrain}) and Afrikaans testing set(F_{Atest})

Category	Symbol	Numeric	Code	English	Afrikaans	F_E	F_{Atrain}	F_{Atest}
Vowels	a	97	a	dug	kat	2.84%	3.74%	3.53%
	e	101	e	fear	lees	1.19%	1.93%	2.22%
	i	105	i	meet	tier	4.59%	5.25%	5.05%
	o	111	o	poor	oop	0.27%	1.70%	1.16%
	u	117	u	boot	soek	0.57%	1.72%	2.53%
	y	121	y		nuut	0.85%	0.79%	0.60%
	:	130	eh:		sê	0.31%	0%	0%
		131	eh	met	met	3.24%	1.69%	2.05%
		132	ao	paw	kos	1.21%	3.06%	2.51%
	:	133	ao:	bore	môre	0.64%	0%	0%
	ø	142	iax		kleur	0.60%	0.66%	0.74%
		143	ax	ago	is	11.18%	9.70%	13.30%
	:	144	ax:	flower	wîe	0.63%	0.01%	0%
	æ	145	ae	bat	ek	2.28%	0.30%	1.02%
	œ	149	oe		nut	2.57%	0.81%	1.17%
	œ:	150	oe:	fur	brûe	0.97%	0%	0%
	:	247	aa	bar	aan	1.51%	2.87%	2.49%
Diphthongs	:i	126	a:i	bite	saai	1.26%	1.12%	0.34%
	o:i	128	o:i		mooi	0.01%	0.36%	0.32%
	oi	134	oi	boy		0.39%	0.08%	0.50%
	i	140	ehi		bedjie	0.05%	0%	0.16%
	i	151	axi	fate	ys	1.47%	1.50%	1.46%
	ui	153	ui		moeite	0.00%	0.65%	0.46%
	iu:	210	iu:	due	leeu	0.05%	0.49%	0.30%
	œu	211	oeu	goat	oud	0.60%	0.46%	0.45%
	œy	217	oey		lui	0.10%	0.86%	0.84%
	õ:	245	aw	brow		0.32%	0%	0%
Nasals	m	109	m	mat	mat	2.77%	2.92%	2.11%
	n	110	n	net	net	7.82%	6.56%	5.69%
		205	ng	sing	sing	1.37%	1.23%	1.56%

Table A.4: Phonetic classes, symbols, Sunspeech labels and computer phonetic labels with English and Afrikaans examples, as well as the percentage occurrence relative to all labels of the label in the English data (F_E) and the Afrikaans training set(F_{Atrain}) and Afrikaans testing set(F_{Atest})

Category	Symbol	Numeric	Code	English	Afrikaans	F_E	F_{Atrain}	F_{Atest}
Fricatives	f	102	f	fat	vars	2.18%	3.39%	2.91%
	h	104	h	hat	huis	0.80%	0.71%	0.11%
	s	115	s	sit	slim	5.67%	6.10%	6.25%
	v	118	v	van	was	1.90%	1.75%	1.72%
	x	120	x		gaan	0.01%	2.65%	2.59%
	z	122	z	zip	soem	1.64%	0.67%	0.53%
	θ	171	th	thin		0.53%	0%	0%
		172	dh	then		1.53%	0%	0%
		188	sh	ship	Sjina	1.48%	0.44%	0.30%
		195	zh	vision	genre	0.36%	0%	0.14%
Affricates	ts^h	181	ts	cats		0.46%	0.06%	0.32%
	dz	184	dz	cads		0.12%	0%	0%
	t^h	191	ch	chin		1.31%	0.68%	0.68%
	d	193	jh	jam		0.82%	0%	0.01%
Liquids	r	114	r	rat		2.95%	4.33%	2.88%
	R	82	r		rooi	0%	4.15%	2.54%
		94	r		berge	0%	0%	0.26%
	l	108	l	lot	lou	2.97%	3.50%	2.97%
		218	/	refers to a flap		0.53%	0.33%	0.58%
Glides	j	106	j	yet	jas	0.47%	1.15%	1.03%
	w	119	w	win	kwes	1.58%	0.10%	0.89%
Stops	b	98	b	bat	bed	1.50%	1.54%	2.93%
	d	100	d	dog	dam	2.36%	3.95%	3.51%
	g	103	g	go	berge	0.66%	0.47%	1.02%
	k	107	k	kit	kar	3.21%	2.79%	3.66%
	p	112	p	pet	pos	2.04%	0.95%	2.47%
	t	116	t	tip	taal	5.85%	4.63%	5.49%
Other		42	sil	silence		3.42%	3.72%	2.12%
	?	63	?	Unknown		1.97%	0.47%	0.42%

that do not have representative examples, but yet occur in the database. The significant examples are listed next, in order of decreasing occurrence.

oi : **toyitoyi**, boikotters, **mooi, rooi**

ts^h : **tsetsevlieg**, maatskappy, Suid-Afrika, **tj**inkeringtjees

t^h : **tj**inkeringtjees, **Charles**, Gorbatsjof, **tj**elloversameling

The set of phonemes used to label the SUN Speech Database represent the union of the phonemes found in Afrikaans and English. The labels were assigned phonetically and there should be close correspondence between data in the two languages with the same phonetic labels. Because such an expanded set of phonemes were used for labelling, many of the phonemes that do not usually appear in English phonetic transcriptions do appear as phone labels in the English transcriptions of this database. This has the advantage that for almost all the phonemes found in Afrikaans there are labelled examples in English, with the exception of the phonemes [R] and [] that do not appear in the English part of the database. The [r], [R] and [] categories are combined into a single [r] category since the [R] and [] categories are not well represented in either of the sets used for training, and because the distinction between the three phones is not important for word recognition purposes. For a few of the phonemes the number of examples that appear in the English transcriptions are very little. They are the phonemes(with the number of occurrences in brackets) : [o:i] (11), [ui] (2) and [x] (8).

Appendix B

TIMIT - SUN Speech phonetic mapping

This appendix (Appendix B) presents almost verbatim work that was performed by Dr. Hendrik Boshoff in his capacity as phonetic expert and is included in this thesis for completeness and because it has not been published elsewhere.

A mapping from SUN Speech symbols to those of TIMIT was required, in order to do cross-database training of phonetic models.

Mapping from one set of phonetic symbols to another is fraught with difficulty, especially when more than one language is involved. Vowels are especially problematic, as dynamic features contribute to subtle differences. In the present case, SUN Speech already contains English and Afrikaans speech, but some problems remain.

A few significant differences in approach between the databases must be mentioned:

- TIMIT views stops as potentially two segments, closure and release. An intervocalic stop of [t] for example, is always transcribed as 'TCL T.' In other positions, the transcription depends upon the actual realization. This allows the affricate [ts] to

be rendered as 'TCL S.' SUN Speech segments all phases of the stop together, and provides separate symbols for all affricates.

- SUN Speech makes provision for front rounded vowels, and when judged appropriate, English vowels are also transcribed using these symbols. This is a somewhat more 'phonetic' approach, versus that of TIMIT, which is more 'phonemic' with respect to vowels.

- TIMIT groups all vocalic sounds together, and does not indicate diphthongization. SUN Speech has an extensive set of diphthongs, and also labels quantity to some extent.

- TIMIT explicitly indicates beginning and end of speech, and sometimes primary and secondary stress. Both these types of transcription are absent in SUN Speech.

It was assumed that every symbol of SUN Speech had to be mapped to one of TIMIT and vice versa. In some cases this was highly artificial, and a 'matching quality' figure was introduced. This ranges from 1 to 3, with the following meanings. 1: The phonemes indicated by the symbols match closely, and some allophones are likely to be identical across the databases. 2: The phonemes are not identical, but are 'neighbours' in phonetic space. 3: The match is poor, but some features are similar, eg place or manner of articulation.

Following are two tables according to the SUN Speech organisation, with the preferred equivalents from TIMIT.

Table B.1: Mapping from SUN Speech to TIMIT symbols (vocoids)

| SUN Speech | | | | | | Match | TIMIT | |
Category	Sym	Num	Code	Eng	Afr	quality	code	word
Vowels	a	97	a	dug	kat	1	ah	but
						2	ax-h	suspect
	e	101	e	fear	lees	3	ey	bait
	i	105	i	meet	tier	1	iy	beet
	o	111	o	poor	oop	2	ow	boat
	u	117	u	boot	soek	1	uw	boot
						2	ux	toot
						2	uh	book
	y	121	y		nuut	3	ux	toot
						3	uw	boot
	:	130	eh:	van	sê	2	eh	bet
		131	eh	met	met	1	eh	bet
		132	ao	paw	kos	1	ao	bought
	:	133	ao:	bore	môre	2	ao	bought
	ø	142	iax		kleur	3	ey	bait
						3	uw	boot
		143	ax	ago	is	1	ax	about
						1	ix	debit
						1	ih	bit
	:	144	ax:	flower	wîe	3	axr	butter
						3	er	bird
	æ	145	ae	bat	ek	1	ae	bat
	œ	149	oe		nut	2	ih	bit
						2	ix	debit
	œ:	150	oe:	fur	brûe	3	er	bird
	:	247	aa	bar	aan	2	aa	bott
Diphthongs	:i	126	a:i	bite	saai	1	ay	bite
	o:i	128	o:i		mooi	3	oy	boy
	oi	134	oi	boy		1	oy	boy
	i	140	ehi		bedjie	3	eh	bet
						3	ey	bait
	i	151	axi	fate	ys	1	ey	bait
	ui	153	ui		moeite	3	ey	bait
						3	ux	toot
	iu:	210	iu:	due	leeu	3	ow	boat
						3	ey	bait
	œu	211	oeu	goat	oud	1	ow	boat
	œy	217	oey		lui	3	ey	bait
						3	oo	toot
	õ:	245	aw	brow		1	aw	bout

Table B.2: Mapping from SUN Speech to TIMIT symbols (contoids)

Category	Sym	Num	Code	Eng	Afr	Match quality	code	word
Nasals	m	109	m	mat	mat	1	m	mom
						1	em	bottom
	n	110	n	net	net	1	n	noon
						1	en	button
		205	ng	sing	sing	1	ng	sing
						1	eng	Washington
Fricatives	f	102	f	fat	vars	1	f	fin
	h	104	h	hat	huis	1	hh	hay
						1	hv	ahead
	s	115	s	sit	slim	1	s	sea
	v	118	v	van	was	1	v	van
	x	120	x		gaan	3	hh	hay
						3	k	key
	z	122	z	zip	soem	1	z	zone
	θ	171	th	thin		1	th	thin
		172	dh	then		1	dh	then
		188	sh	ship	Sjina	1	sh	she
		195	zh	vision	genre	1	zh	azure
Affricates	ts^h	181	ts	cats		3	t	tea
						3	s	sea
	dz	184	dz	cads		3	d	day
						3	z	zone
	t^h	191	ch	chin		1	ch	choke
	d	193	jh	jam		1	jh	joke
Glides	j	106	j	yet	jas	1	y	yacht
	w	119	w	win	kwes	1	w	way
Liquids	r	114	r	rat		1	r	ray
	R	82	r		rooi	2	r	ray
		94	r		berge	2	r	ray
	l	108	l	lot	lou	1	l	lay
						1	el	bottle
		218	/	(flap)		1	dx	muddy
						2	nx	winner
Stops	b	98	b	bat	bed	1	b	bee
	d	100	d	dog	dam	1	d	day
	g	103	g	go	berge	1	g	gay
	k	107	k	kit	kar	1	k	kite
						3	q	bat
	p	112	p	pet	pos	1	p	pea
	t	116	t	tip	taal	1	t	tea

Appendix C

MCE update derivations

C.1   Mixture weight derivative

$$\frac{\partial}{\partial \bar{c}_{jk}^{(i)}} \log b_j^{(i)}(\mathbf{x}_t) = (b_j^{(i)}(\mathbf{x}_t))^{-1} \frac{\partial b_j^{(i)}(\mathbf{x}_t)}{\partial \bar{c}_{jk}^{(i)}} \tag{C.1}$$

$$= (b_j^{(i)}(\mathbf{x}_t))^{-1} \sum_{k'}^{M} \frac{\partial b_j^{(i)}(\mathbf{x}_t)}{\partial c_{jk'}^{(i)}} \frac{\partial c_{jk'}^{(i)}}{\partial \bar{c}_{jk}^{(i)}} \tag{C.2}$$

$$= \sum_{k'}^{M} \frac{\mathcal{N}[\mathbf{x}_t, \boldsymbol{\mu}_{jk'}^{(i)} \Sigma_{jk'}^{(i)}]}{b_j^{(i)}(\mathbf{x}_t)} \frac{\partial c_{jk'}^{(i)}}{\partial \bar{c}_{jk}^{(i)}} \tag{C.3}$$

where

$$\frac{\partial c_{jk'}^{(i)}}{\partial \bar{c}_{jk}^{(i)}} = \frac{\partial}{\partial \bar{c}_{jk}^{(i)}} \left[\frac{e^{\bar{c}_{jk'}^{(i)}}}{\sum_l e^{\bar{c}_{jl}^{(i)}}} \right] \tag{C.4}$$

$$= \delta(k = k') \frac{\partial}{\partial \bar{c}_{jk}^{(i)}} \left[\frac{e^{\bar{c}_{jk}^{(i)}}}{\sum_l e^{\bar{c}_{jl}^{(i)}}} \right] + \delta(k \neq k') \frac{\partial}{\partial \bar{c}_{jk}^{(i)}} \left[\frac{e^{\bar{c}_{jk'}^{(i)}}}{\sum_l e^{\bar{c}_{jl}^{(i)}}} \right] \tag{C.5}$$

$$= \delta(k = k') \frac{e^{\bar{c}_{jk}^{(i)}} \sum_l e^{\bar{c}_{jl}^{(i)}} - e^{\bar{c}_{jk}^{(i)}} e^{\bar{c}_{jk}^{(i)}}}{(\sum_l e^{\bar{c}_{jl}^{(i)}})^2} + \delta(k \neq k') \frac{-e^{\bar{c}_{jk'}^{(i)}} e^{\bar{c}_{jk}^{(i)}}}{(\sum_l e^{\bar{c}_{jl}^{(i)}})^2} \tag{C.6}$$

$$= \delta(k = k') \frac{e^{\bar{c}_{jk}^{(i)}}}{\sum_l e^{\bar{c}_{jl}^{(i)}}} \frac{\sum_{l \neq k} e^{\bar{c}_{jl}^{(i)}}}{\sum_l e^{\bar{c}_{jl}^{(i)}}} + \delta(k \neq k') c_{jk'} c_{jk} \tag{C.7}$$

$$= \delta(k = k') c_{jk}(1 - c_{jk}) + \delta(k \neq k') c_{jk'} c_{jk} \tag{C.8}$$

$$= c_{jk}^{(i)} \delta(k' - k) - c_{jk}^{(i)} c_{jk'}^{(i)}. \tag{C.9}$$

and therefore

$$\frac{\partial}{\partial \bar{c}_{jk}^{(i)}} \log b_j^{(i)}(\mathbf{x}_t) = \sum_{k'}^{M} \frac{\mathcal{N}[\mathbf{x}_t, \boldsymbol{\mu}_{jk'}^{(i)}, \Sigma_{jk'}^{(i)}]}{b_j^{(i)}(\mathbf{x}_t)} \left[c_{jk}^{(i)} \delta(k' - k) - c_{jk}^{(i)} c_{jk'}^{(i)} \right] \tag{C.10}$$

$$= c_{jk}^{(i)} \sum_{k'}^{M} \frac{\mathcal{N}[\mathbf{x}_t, \boldsymbol{\mu}_{jk'}^{(i)}, \Sigma_{jk'}^{(i)}]}{b_j^{(i)}(\mathbf{x}_t)} \left[\delta(k' - k) - c_{jk'}^{(i)} \right] \tag{C.11}$$

$$= \frac{c_{jk}^{(i)}}{b_j^{(i)}(\mathbf{x}_t)} \left[\mathcal{N}[\mathbf{x}_t, \boldsymbol{\mu}_{jk}^{(i)}, \Sigma_{jk}^{(i)}] - \sum_{k'}^{M} c_{jk'}^{(i)} \mathcal{N}[\mathbf{x}_t, \boldsymbol{\mu}_{jk'}^{(i)}, \Sigma_{jk'}^{(i)}] \right] \tag{C.12}$$

$$= c_{jk}^{(i)} \left[\frac{\mathcal{N}[\mathbf{x}_t, \boldsymbol{\mu}_{jk}^{(i)}, \Sigma_{jk}^{(i)}]}{b_j^{(i)}(\mathbf{x}_t)} - 1 \right]. \tag{C.13}$$

C.2 Transition probability derivative

$$\frac{\partial}{\partial \bar{a}_{jj'}^{(i)}} g_i(X; \Lambda) = \sum_{t=1}^{T(X)} \sum_{s=1}^{N} \delta(\bar{q}_{t-1} - j) \delta(\bar{q}_t - s) \frac{\partial}{\partial \bar{a}_{jj'}^{(i)}} \log a_{js}^{(i)} \tag{C.14}$$

$$= \sum_{t=1}^{T(X)} \sum_{s=1}^{N} \delta(\bar{q}_{t-1} - j) \delta(\bar{q}_t - s) \frac{1}{a_{js}^{(i)}} \frac{\partial a_{js}^{(i)}}{\partial \bar{a}_{jj'}^{(i)}} \tag{C.15}$$

where the derivative of $a_{js}^{(i)}$ with respect to $\bar{a}_{jj'}^{(i)}$ is similar to the derivation in Equations C.4-C.9, giving

$$\frac{\partial a_{js}^{(i)}}{\partial \bar{a}_{jj'}^{(i)}} = a_{jj'}^{(i)} \delta(j' - s) - a_{jj'}^{(i)} a_{js}^{(i)} \tag{C.16}$$

and therefore

$$\frac{\partial}{\partial \bar{a}_{jj'}^{(i)}} \log g_i(X; \Lambda) = \sum_{t=1}^{T(X)} \sum_{s=1}^{N} \delta(\bar{q}_{t-1} - j)\delta(\bar{q}_t - s)\frac{a_{jj'}^{(i)}\delta(j' - s) - a_{jj'}^{(i)}a_{js}^{(i)}}{a_{js}^{(i)}} \tag{C.17}$$

$$= \sum_{t=1}^{T(X)} \sum_{s=1}^{N} \delta(\bar{q}_{t-1} - j)\delta(\bar{q}_t - s)[\delta(j' - s) - a_{jj'}^{(i)}]. \tag{C.18}$$

Bibliography

[1] L. Rabiner, "An introduction to hidden Markov models," *IEEE Signal Processing Magazine*, vol. 5, pp. 4–16, Jan. 1988.

[2] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, vol. 13, pp. 45–57, Sep. 1996.

[3] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," in *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp. 100–109, Feb. 1986.

[4] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proc. ICASSP '92*, (San Francisco, CA), March 1992.

[5] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. DARPA Speech Natural Language Workshop*, pp. 357–362, Feb. 1992.

[6] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura, and Y. Sagisaka, "Japanese speech databases for robust speech recognition," in *Proc. ICSLP '96*, Vol. 4, (Philadelphia, PA), pp. 2199–2202, Oct. 1996.

[7] J.-L. Gauvain, L. F. Lamel, and M. Eskénazi, "Design considerations and text selection for BREF, a large French read-speech corpus," in *Proc. ICSLP '90*, (Kobe, Japan), pp. 1097–1100, Nov. 1990.

[8] D. Langmann, R. Haeb-Umbach, L. Boves, and E. den Os, "FRESCO: The French telephone speech data collection - part of the European SpeechDat(m) project," in *Proc. ICSLP '96*, Vol. 3, (Philadelphia, PA), pp. 1918–1921, Oct. 1996.

[9] K. J. Kohler, "Labelled data bank of spoken standard German: The Kiel corpus of read/spontaneous speech," in *Proc. ICSLP '96*, Vol. 3, (Philadelphia, PA), pp. 1938–1941, Oct. 1996.

[10] T. Bub and J. Schwinn, "VERBMOBIL: The evolution of a complex large speech-to-speech translation system," in *Proc. ICSLP '96*, Vol. 4, (Philadelphia, PA), pp. 2371–2374, Oct. 1996.

[11] H. Höge, C. Draxler, H. van den Heuvel, F. Johansen, E. Sanders, and H. Tropf, "SpeechDat multilingual speech databases for teleservices: Across the finish line," in *Proc. Eurospeech '99*, (Budapest, Hungary), pp. 2699–2702, Sep. 1999.

[12] T. Waardenburg, J. du Preez, and M. Coetzer, "The automatic recognition of stop consonants using hidden Markov models," in *Proc. ICASSP '92*, (San Francisco, CA), pp. 1585–1588, March 1992.

[13] F. de Wet and E. Botha, "Towards speech technology for South African languages: Automatic speech recognition in Xhosa," Accepted for publication in the *South African Journal of African Languages*, Mar. 1999.

[14] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken-language understanding in the MIT Voyager system," *Speech Communication*, vol. 17, pp. 1–18, Aug. 1995.

[15] J. Billa, K. Ma, J. W. McDonough, G. Zavaliagkos, D. R. Miller, K. N. Ross, and A. El-Jaroudi, "Multilingual speech recognition: The 1996 Byblos callhome system," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 363–366, Sep. 1997.

[16] B. Wheatley, K. Kondo, W. Anderson, , and Y. Muthusamy, "An evaluation of cross-language adaptation for rapid HMM development in a new language," in *Proc. ICASSP '94*, (Adelaide, Australia), pp. I–237 – I–240, Apr. 1994.

[17] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 371–374, Sep. 1997.

[18] C. Wang, J. Glass, H. Meng, J. Polifroni, S. Seneff, and V. Zue, "Yinhe: A Mandarin Chinese version of the Galaxy system," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 351–354, Sep. 1997.

[19] F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, "A study of multilingual speech recognition," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 359–362, Sep. 1997.

[20] P. Bonaventura, F. Gallocchio, and G. Micca, "Multilingual speech recognition for flexible vocabularies," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 355–358, Sep. 1997.

[21] J. Köhler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *Proc. ICASSP '98*, (Seattle, USA), pp. 417 – 420, May 1998.

[22] U. Bub, J. Köhler, and B. Imperl, "In-service adaptation of multilingual hidden Markov models," in *Proc. ICASSP '97*, (Munich, Germany), pp. 1451 – 1454, Apr. 1997.

[23] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 35, pp. 751–763, June 1987.

[24] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 806–841, Apr. 1991.

[25] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York, NY: John Wiley & Sons, 1973.

[26] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[27] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, Apr. 1995.

[28] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.

[29] T. Matsui and S. Furui, "A study of speaker adaptation based on minimum classification error training," in *Proc. Eurospeech '95*, (Madrid, Spain), pp. 81–84, Sep. 1995.

[30] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. ICSLP '98*, Vol. 5, (Sydney, Australia), pp. 1819–1822, Nov. 1998.

[31] ARPA, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," NIST Speech Disc CD1-1.1, Dec. 1990.

[32] C. Nieuwoudt and E. Botha, "Cross-language use of acoustic information for automatic speech recognition," Submitted to *Speech Communication*, May 2000.

[33] C. Nieuwoudt and E. Botha, "Adaptation of acoustic models for multilingual recognition," in *Proc. Eurospeech '99*, (Budapest, Hungary), pp. 907–910, Sep. 1999.

[34] C. Nieuwoudt and E. Botha, "Multilingual training of acoustic models in automatic speech recognition," Accepted for publication in the *South African Computer Journal*, Oct. 1999.

[35] C. Nieuwoudt and E. Botha, "Connected digit recognition in Afrikaans using hidden Markov models," *South African Computer Journal*, pp. 85–91, Jul. 1999.

[36] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition using hidden Markov models," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-37, no. ASSP-37, p. 1214, 1989.

[37] Y. Gotoh, M. M. Hochberg, and H. F. Silverman, "Using MAP estimated parameters to improve HMM speech recognition performance," in *Proc. ICASSP '94*, (Adelaide, Australia), pp. I–229 – I–232, Apr. 1994.

[38] X. Wang, *Incorporating Knowledge on Segmental Duration in HMM-Based Continuous Speech Recognition.* PhD thesis, University of Amsterdam, Amsterdam, The Netherlands, Apr. 1997.

[39] J. Du Preez, "Modelling durations in hidden Markov models with application to word spotting," in *Proc. IEEE South African symposium on Communications and Signal Processing*, (Fourways, South Africa), pp. 1–5, Aug. 1991.

[40] D. Burshtein, "Robust parametric modeling of durations in hidden Markov models," in *Proc. ICASSP '95*, (Detroit, MI), pp. 548 – 551, May 1995.

[41] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 39, pp. 1–38, 1977.

[42] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Information Theory*, vol. IT-13, pp. 260–269, Apr. 1967.

[43] F. K. Soong and E. Huang, "A tree-trellis based fast search for finding the N best sentence hypotheses in continuous speech recognition," in *Proc. ICASSP '91*, (Toronto, Canada), pp. 705–708, May 1991.

[44] M. Zissman, "Comparision of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 31–44, Jan. 1996.

[45] E. Barnard and Y. Yan, "Toward new language adaptation for language identification," *Speech Communication*, vol. 21, pp. 245–254, 1997.

[46] A. Lavie, A. Waibel, L. Levin, D. Gates, M. Gavalda, T. Zeppenfeld, P. Zhan, and O. Glickman, "Translation of conversational speech with JANUS-II," in *Proc. ICSLP '96*, Vol. 4, (Philadelphia, PA), pp. 2375–2378, Oct. 1996.

[47] T. Schultz, D. Koll, and A. Waibel, "Japanese LVCSR on the spontaneous scheduling task with JANUS-3," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 367–370, Sep. 1997.

[48] J. Köhler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. ICSLP '96*, Vol. 4, (Philadelphia, PA), pp. 2195–2198, Oct. 1996.

[49] U. Uebler, M. Schussler, and H. Niemann, "Bilingual and dialectal adaptation and retraining," in *Proc. ICSLP '98*, Vol. 5, (Sydney, Australia), pp. 1815–1818, Nov. 1998.

[50] X. Huang and K.-F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," in *Proc. ICASSP '91*, (Toronto, Canada), pp. 877–880, May 1991.

[51] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," *Speech Communication*, vol. 5, pp. 183–197, 1986.

[52] C. Leggetter, *Improved acoustic modelling for HMMs using linear transformations.* PhD thesis, Cambridge University, 1995.

[53] H. Wakita, "Normalisation of vowels by vocal tract length and its application to vowel identification," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 25, no. 25, pp. 183–192, 1977.

[54] H. C. Choi and R. W. King, "On the use of spectral transformation for speaker adaptation in HMM based isolated-word speech recognition," *Speech Communication*, vol. 17, pp. 131–144, Aug. 1995.

[55] L. Uebel and P. Woodland, "An investigation into vocal tract length normalisation," in *Proc. Eurospeech '99*, (Budapest, Hungary), pp. 2527–2530, Sep. 1999.

[56] Y. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 380–394, July 1994.

[57] M. H. DeGroot, *Probability and Statistics.* Reading, MA: Addison-Wesley, 1975.

[58] J.-L. Gauvain and C. Lee, "Bayesian learning for hidden Markov models with Gaussian mixture state observation densities," *Speech Communication*, vol. , pp. 205–213, June 1992.

[59] M. H. DeGroot, *Optimal Statistical Decisions.* New York, NY: McGraw-Hill, 1970.

[60] C.-H. Lee and J.-L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," in *Proc. ICASSP '93*, (Minneapolis, MN), pp. II–558–II–561, Apr. 1993.

[61] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

[62] L. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Information Theory*, vol. IT-28, pp. 729–734, Sep. 1982.

[63] C. Leggetter and P. Woodland, "Speaker adaptation of HMMs using linear regression," Tech. Report CUED/F-INFENG/TR. 181, Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, England, June 1994.

[64] H. Matsukoto and H. Inoue, "A piecewise linear spectral mapping for supervised speaker adaptation," in *Proc. ICASSP '92*, (San Francisco, CA), pp. 449–452, March 1992.

[65] A. Sankar, L. Neumeyer, and M. Weintraub, "An experimental study of acoustic adaptation algorithms," in *Proc. ICASSP '95*, (Detroit, MI), pp. 713 – 716, May 1995.

[66] V. Nagesha and L. Gillick, "Studies in transformation-based adaptation," in *Proc. ICASSP '97*, (Munich, Germany), pp. 1031 – 1034, Apr. 1997.

[67] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 71–77, Jan. 1998.

[68] A. J. Hewett, *Training and speaker adaptation in template-based speech recognition.* PhD thesis, Cambridge University, 1989.

[69] S. Cox, "A speaker adaptation technique using linear regression," in *Proc. ICASSP '95*, (Detroit, MI), pp. 700 – 703, May 1995.

[70] C. Leggetter and P. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. ARPA Spoken Language Technology Workshop*, (Barton Creek), 1995.

[71] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, Sep. 1995.

[72] M. Gales and P. Woodland, "Mean and variance adaptation with the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, Oct. 1996.

[73] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, Apr. 1998.

[74] V. Abrash, A. Sankar, H. Franco, and M. Cohen, "Acoustic adaptation using nonlinear transformations of HMM parameters," in *Proc. ICASSP '96*, (Atlanta, GA), pp. 729 – 732, May 1996.

[75] J. R. Bellegarda, P. V. de Souza, A. J. Nádas, D. Nahamoo, M. A. Picheny, and L. R. Bahl, "Robust speaker adapation using a piecewise linear acoustic mapping," in *Proc. ICASSP '92*, (San Francisco, CA), pp. 445–448, March 1992.

[76] J. R. Bellegarda, P. V. de Souza, D. Nahamoo, M. Padmanabhan, M. A. Picheny, and L. R. Bahl, "Experiments using data augmentation for speaker adaptation," in *Proc. ICASSP '95*, (Detroit, MI), pp. 692 – 695, May 1995.

[77] J. Ishii and M. Tonomura, "Speaker normalization and adaptation based on linear transformation," in *Proc. ICASSP '97*, (Munich, Germany), pp. 1055 – 1058, Apr. 1997.

[78] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," in *Proc. ICASSP '95*, (Detroit, MI), pp. 680–683, May 1995.

[79] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 294–300, July 1996.

[80] E. Thelen, X. Aubert, and P. Beyerlein, "Speaker adaptation in the Philips system for large vocabulary continuous speech recognition," in *Proc. ICASSP '97*, (Munich, Germany), pp. 1035 – 1038, Apr. 1997.

[81] W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors," in *Proc. Eurospeech '99*, (Budapest, Hungary), pp. 1–4, Sep. 1999.

[82] D.E.Rumelhart, G.E.Hinton, and R.J.Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[83] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-31, no. ASSP-31, p. 814, 1983.

[84] H. Franco and A. Serralheiro, "Training HMMs using a minimum recognition approach," in *Proc. ICASSP '91*, (Toronto, Canada), pp. 357–360, May 1991.

[85] J.-K. Chen and F. Soong, "An N-best candidates-based discriminative training for speech recognition applications," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 206–216, Jan. 1994.

[86] E. McDermott, *Discriminative Training for Speech Recognition*. PhD thesis, Waseda University, Japan, March 1997.

[87] A. Nadas, D. Nahamoo, and M. A. Picheny, "On a model-robust training method for speech recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-36, no. ASSP-36, p. 1432, 1988.

[88] R. Cardin, Y. Normandin, and R. D. Mori, "High performance connected digit recognition using maximal mutual information estimation," in *Proc. ICASSP '91*, (Toronto, Canada), pp. 533–536, May 1991.

[89] Y. Normandin and S. Morgera, "An improved MMIE training algorithm for speaker-independent, small vocabulary, continuous speech recognition," in *Proc. ICASSP '91*, (Toronto, Canada), pp. 537–540, May 1991.

[90] S. Kapadia, V. Valtchev, and S. J. Young, "MMI training for continuous phoneme recognition on the TIMIT database," in *Proc. ICASSP '93*, (Minneapolis, MN), Apr. 1993.

[91] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, pp. 3043–3054, December 1992.

[92] W. Chou, B.-H. Juang, and C.-H. Lee, "Segmental GPD training of HMM based speech recognizer," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, (San Francisco), IEEE, pp. 473–476, 1992.

[93] O. Kwon and C. Un, "Performance of HMM-based speech recognizers with discriminative state-weights," *Speech Communication*, vol. 19, pp. 197–205, 1996.

[94] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum error rate training based on N-best string models," in *Proc. ICASSP '93*, (Minneapolis, MN), pp. I–652 – I–655, Apr. 1993.

[95] E. McDermott and S. Katagiri, "String-level MCE for continuous phoneme recognition," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 123–126, Sep. 1997.

[96] P.-C. Chang, S.-H. Chen, and B.-H. Juang, "Discriminative analysis of distortion sequences in speech recognition," in *Proc. ICASSP '91*, (Toronto, Canada), pp. 549–552, May 1991.

[97] P.-C. Chang and B.-H. Juang, "Discriminative template training for dynamic programming speech recognition," in *Proc. ICASSP '92*, (San Francisco, CA), pp. 493–496, March 1992.

[98] T. Komori and S. Katagiri, "Application of a generalized probabilistic descent method to dynamic time warping-based speech recognition," in *Proc. ICASSP '92*, (San Francisco, CA), pp. 497–500, March 1992.

[99] E. McDermott and S. Katagiri, "Prototype-based MCE/GPD training for word spotting and connected word recognition," in *Proc. ICASSP '93*, (Minneapolis, MN), pp. II–291 – II–294, Apr. 1993.

[100] T. Komori and S. Katagiri, "An optimal learning method for minimizing spotting errors," in *Proc. ICASSP '93*, (Minneapolis, MN), pp. II–271 – II–274, Apr. 1993.

[101] R. Sukkar and J. Wilpon, "A two pass classifier for utterance rejection in keyword spotting," in *Proc. ICASSP '93*, (Minneapolis, MN), pp. II–451 – II–454, Apr. 1993.

[102] R. A. Sukkar, "Rejection for connected digit recognition based on GPD segmental discrimination," in *Proc. ICASSP '94*, (Adelaide, Australia), pp. I–393 – I–396, Apr. 1994.

[103] A. Biem and S. Katagiri, "Feature extraction based on minimum classification error/generalized probabilistic descent method," in *Proc. ICASSP '93*, (Minneapolis, MN), pp. II–275 – II–278, Apr. 1993.

[104] A. Biem and S. Katagiri, "Filter bank design based on discriminative feature extraction," in *Proc. ICASSP '94*, (Adelaide, Australia), pp. I–485 – I–488, Apr. 1994.

[105] C. Rathinavelu and L. Deng, "Use of generalized dynamic feature parameters for speech recognition: Maximum likelihood and minimum classification error approaches," in *Proc. ICASSP '95*, (Detroit, MI), pp. 373 – 376, May 1995.

[106] H. Watanabe, T. Yamaguchi, and S. Katagiri, "Discriminative metric design for pattern recognition," in *Proc. ICASSP '95*, (Detroit, MI), pp. 3439 – 3442, May 1995.

[107] C. Rathinavelu and L. Deng, "HMM-based speech recognition using state-dependent linear transforms on mel-warped DFT features," in *Proc. ICASSP '96*, (Atlanta, GA), pp. 9 – 12, May 1996.

[108] E. McDermott, E. Woudenberg, and S. Katagiri, "A telephone-based directory assistance system adaptively trained using minimum classification error/generalised probabilistic descent," in *Proc. ICASSP '96*, (Atlanta, GA), pp. 3347 – 3350, May 1996.

[109] K. Laurila, M. Vasilache, and O. Viikki, "A combination of discriminative and maximum likelihood techniques for noise robust speech recognition," in *Proc. ICASSP '98*, (Seattle, USA), pp. 85 – 88, May 1998.

[110] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary speech recognition," in *Proc. ICASSP '92*, (San Francisco, CA), pp. I–13–I–16, March 1992.

[111] C. Rathinavelu, "Minimum classification error linear regression (MCELR) for speaker adaptation using HMM with trend functions," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 2343–2346, Sep. 1997.

[112] D. B. Paul, "On the interaction between between true source, training and testing language models," in *Proc. ICASSP '91*, (Toronto, Canada), pp. 569–572, May 1991.

[113] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," in *Proc. ICSLP '96*, Vol. 4, (Philadelphia, PA), pp. 2005–2008, Oct. 1996.

[114] V. Diakoloukas and V. Digalakis, "Maximum-likelihood stochachastic-transformation adaptation of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 177–187, March 1999.