MODELING REDUCED PRONUNCIATIONS IN GERMAN

Martine Adda-Decker & Lori Lamel

Spoken Language Processing Group, LIMSI-CNRS, Orsay, France
{madda,lamel}@limsi.fr, http://www.limsi.fr/TLP

Abstract

This paper deals with pronunciation modeling for automatic speech recognition in German with a special focus on reduced pronunciations. Starting with our 65k full form pronunciation dictionary we have experimented with different phone sets for pronunciation modeling. For each phone set, different lexica have been derived using mapping rules for unstressed syllables, where schwa-vowel+[l n m] are replaced by syllabic [l n m]. The different pronunciation dictionaries are used both for acoustic model training and during recognition. Speech corpora consisted of television programmes, which contain signal segments of a varying acoustic and linguistic nature. The speech is produced by a wide variety of speakers, with linguistic styles ranging from prepared to spontaneous speech and with changing background and channel conditions. Experiments were carried out using 4 news programmes and documentaries lasting more than 15 minutes each (with a total of 1h20min). Word error rates obtained vary between 19 and 29%, depending on the programme and the system configuration. Only small differences in recognition rates were measured for the different experimental setups, with slightly better results obtained by the reduced lexica.

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, 145-159.

1. Introduction

Pronunciation variant modeling for automatic speech recognition is a research domain which has attracted much interest in recent years (Rolduc, 1998; SpeechCom, 1999). In previous work (Adda & Lamel, 1999), we have investigated the use of pronunciation variants in speech alignment experiments, where the acoustic score alone drives the aligned pronunciation choice. These experiments were run for English and French. In the following work, we investigate the use of reduced pronunciations during recognition experiments in German. Our first German speech recognition system was developed within the European LE-SQALE project on read newspaper texts (Young et al., 1997; Lamel et al., 1995; Adda-Decker et al., 1996) more than five years ago. In the present contribution, we report on our ongoing work in German speech recognition on broadcast speech with a focus on acoustic modeling and pronunciation variants. Part of this work is funded by the European LE-OLIVE project.

The aim of our study is to investigate the acoustic modeling of reduction phenomena and their impact on speech recognition. In German, long words with complex syllable structures can commonly be observed. Concatenations of complex syllables may result in sequences of 5, 6 and even 7 consonants (e.g. `se`**`lbst-kr`**`itisch`, `Ausku`**`nfts-pfl`**`icht`) in a canonical pronunciation. Such consonant clusters may be subject to more or less severe reductions. Reduction phenomena also concern common words (e.g. `haben` → `ham`, `ein` → `'n`) and numbers (`neunundneunzig` → `neu'neunzig`) where the missing acoustic information is supplied by the higher levels. Unstressed word endings (`könn`**`en`**, `zwisch`**`en`**, `dies`**`em`**...), generally predictable from the syntactic or semantic context, are often loosely articulated and reduced. We may expect that reduction phenomena are less prone to error within words than at word boundaries, where a large number of successor phones are possible. This motivates our experiments in word-final reduction modeling. In this contribution, we start by evaluating different phone sets for pronunciation modeling. Then comparative experiments are carried out using different types of variants, with a special focus on word- or morpheme-final unstressed syllables /ən, əm, əl/. In Section 2, we describe the phone sets used and the different types of pronunciation dictionaries. In Section 3, we give a summary of the acoustic data and the text material used for model estimation. Section 4 gives a brief overview of the transcription system including the automatic acoustic data partitioning, the acoustic phone models, the language models and the decoder. In Section 5, experimental results are presented and discussed.

Table 1. IPA and LIMSI phone set for German (52 vowels and consonants). Symbols for which no *comment* is given are included in all the different phone sets.

IPA	**LIMSI**	***comment***	**example**	**IPA**	**LIMSI**	***comment***	**example**
iː	!	∉ **47**set	viel	p	p		paar
i	i		vital	b	b		bald
ɪ	I		will	t	t		tun
eː	6	∉ **46**set	wen	d	d		doch
e	e		methodisch	k	k		kurz
ɛː	9		gähnen	g	g		gar
ɛ	E		wenn	ʔ	?	*not used*	_ach
ɑ	a		wahr	m	m		man
a	A		man	n	n		noch
oː	0	∉ **47**set	so	ŋ	G		bang
o	o		sofort	f	f		fort
ɔ	O		von	v	v		wann
uː	V	∉ **47**set	zu	s	s		es
u	u		zuvor	z	z		so
ʊ	U		durch	ʃ	S		schön
yː	7	∉ **47**set	müde	ʒ	Z		Genie
y	y		mythologisch	ç	J		ich
ʏ	Y		mündlich	x	K		ach
ø	@		rötlich	h	h		hier
œ	x		örtlich	r	r		rot
ə	X		eine	l	l		los
aj	Q		heim	j	j		ja
aw	q		laut	m	M	∉ **original**	einem
ɔj	c		heute	ən	N	∉ **original**	gehen
a̯	4		für	əl	L	∉ **original**	mittel
i̯	1		Aktion	ər	R		einer

2. Phone sets and pronunciation dictionaries

2.1. *Phone sets for pronunciations and acoustic modeling*

The total phone set used for pronunciations is based on **52** phone symbols (see Table 1) including the 3 syllabic /ən, əm, əl/ symbols (the latter are not in our original pronunciation dictionary). But different phone sets are possible. In particular, consistency in the pronunciation dictionary is easier to achieve with smaller sets. The glottal stop, while generated by the grapheme-phoneme converter is not kept for acoustic modeling in the experiments reported here. Thus the largest phone set used for the acoustic models includes **51** phone symbols plus 3 additional symbols for silence, breath and a filler noise. We experimented with a smaller phone set of **47** phone symbols by removing the distinction between tense vowels (**/i, u, y, o/**) according to whether they carry primary stress or not (duration diacritic). In a **46** phone symbol set the same type of distinction was removed for the tense **/e/** vowel. We trained distinct acoustic models for all the different phone symbol sets.

2.2. *Pronunciation dictionaries*

The pronunciations are derived from a grapheme-to-phoneme converter developed at LIMSI. It is a PERL script including about 350 rules for standard German words, most common German exceptions, foreign characters and most common foreign words. This letter-to-sound converter has been used to build the 65k pronunciation dictionary of our German transcription system. Manual verification has been carried out, where we used the **Duden** Ausspracheworterbuch (Duden, 1990) as a reference. A large majority of the corrected errors are due to unknown morpheme boundaries and to foreign words. The conclusion drawn from this work is that German letter-to-sound conversion is rather straightforward provided the morphological boundaries are known. Alternative pronunciations are added for frequent words when this is deemed appropriate. Pronunciation variants are often needed for frequent words that are subject to reduction (due to poor articulation) or for foreign words that may be pronounced with a more or less close approximation to the rules of the native language.

Some example entries from our **original** pronunciation dictionaries are shown in Table 2. The original full form lexicon contains a very limited number of variants: about 3% of the words have pronunciation variants (lower part of Table 2). These variants have been introduced to describe alternative pronunciations observed for frequent words and proper names. For example the article der has a standard pronunciation /de4/ and a reduced pronunciation /dR/. When automatically aligning speech corpora the standard form /de4/ is preferred for a majority of 65%, the remaining 35% of the utterances are aligned with the reduced /dR/ form. The proper name Peter was aligned with the standard German pronunciation, except for 2% of the utterances where the English form was preferred.

Table 2. Example lexical entries of the **original** pronunciation lexicon. The lower part of the table lists some of the variants in this lexicon.

Achtelfinale	?AKtXlfinalX
Bilanzpressekonferenz	bilAntsprEsXkOnfXrEnts
Einwanderungsbehörde	?QnvAndXrUGsbXh@4dX
Goetheplatz	g@tXplAts
Immobiliengesellschaften	?Imob!l1XngXzElSAftXn
aktuellem	?AktUElXm
der	de4 dR
zwanzig	tsvAntsIJ tsvAntsIk
Anerkennung	?AnRkEnUG ?An?ErkEnUG
Israel	?IsrAel ?IsrAEl
Peter	p6tR p!tR

We have experimented with different pronunciation lexica. Starting with the 65k full form pronunciation dictionary (**original**[1]), different lexica were derived using mapping rules. According to the rules applied here schwa-vowel+[l n m] are replaced by syllabic [l n m] if they occur in word final position or if followed by a consonant. The mapping sequences may be either simply replaced, resulting in the **reduced** lexicon, or added to **optional**ly allow for full or reduced pronunciations. Some examples are given in Table 3 for each of these 3 lexicon types. For each lexicon type the possible phone sets are specified in the right-hand column of Table 3. The **51**, **47**, **46** phone sets include the syllabic [l n m] symbols, the phone sets of size **48**, **44**, **43**

[1] The glottal stop has been removed for these experiments.

do not. For each of the possible combinations of phone sets and pronunciation lexicon types, distinct acoustic phone models have been trained and used during recognition.

Table 3. Example lexical entries with different pronunciations depending on the lexica (original, reduced, optional). The right-hand column indicates the different phone set sizes (#phones) and the list of phones removed from the set of 52 symbols.

lex.	**lexical entry**	**pronunciations**		**#phones (removed)**
orig.	zwischen	tsvISXn		48 (?, N, M, L)
	Achtelfinale	AKtXlfinalX		44 (?, N, M, L, i:, u:, y:, o:)
	aktuellem	AktUElXm		43 (?, N, M, L, i:, u:, y:, o:, e:)
red.	zwischen	tsvISN		51 (?)
	Achtelfinale	AKtLfinalX		47 (?, i:, u:, y:, o:)
	aktuellem	AktUElM		46 (?, i:, u:, y:, o: e:)
opt.	zwischen	tsvISXn	tsvISN	51 (?)
	Achtelfinale	AKtXlfinalX	AKtLfinalX	47 (?, i:, u:, y:, o:)
	aktuellem	AktUElXm	AktUElM	46 (?, i:, u:, y:, o: e:)

3. Speech and text corpora

In this section, we describe the speech corpora used for acoustic model training and for testing, as well as the written text material from which the system's vocabulary has been selected and language models have been estimated.

3.1. *Broadcast speech data*

Acoustic models were estimated from audio data from ARTE (a bilingual French-German TV station). This data was extracted from the ARTE programming of the last four years according to ARTE's interests (social, cultural or political issues). About 20 hours of transcribed (Barras et al., 1998) German TV broadcasts (news and documents) were used for training.

Four files (2 news broadcasts, 2 documentaries), totalling 1 hour and 20 minutes of audio data, were used for testing (see Table 4). Documentary files contain a single audio document each, whereas the news files contain a collection of several news items.

Table 4. Test data description

show	# sentences	# words	duration
news:			
arte_97:01:29-30	334	2512	15'
arte_97:01:20-23	545	4950	24'
documentaries:			
arte_98:09:28	344	2622	15'
arte_99:02:09	554	2683	20'

3.2. *Text and transcript data*

Written language material is used for vocabulary selection and language model training. Most of the written data come from newspaper texts, but audio transcripts, even if only limited amounts are available, have proved to be very helpful for vocabulary and language model development. About 200k words of audio data transcripts have been added to the German text corpora. These text corpora include different sources. Among the most important we can cite the following: **Deutsche Presse Agentur (German Press Agency)** with about 30M words (years 1993-1996, distributed by the LDC); **Frankfurter Rundschau** newspaper text (about 35 M words) from the ECI (European Corpus Initiative); Berliner TAgesZeitung (**TAZ**) with about 150 M words (years 1986-99) purchased directly from the newspaper; **Die Welt**, years 1996-98, including 20 M words obtained via the Web.

The text data need to be preprocessed for lexicon and language model (LM) development. The different text sources are gathered in different formats with different mark-ups. Therefore each source requires different manipulations. Once the roughly cleaned texts are available, further normalization and processing is needed to prepare them for word list selection and language modeling. The motivation for normalization is to reduce lexical variability so as to increase the coverage for a fixed size task vocabulary. We have chosen to maintain case distinction for German in the vocabulary and language modeling. Recognition error rates, however, are currently computed without case distinction.

4. System description

Our broadcast transcription system comprises mainly two major processing procedures: the data partitioning, which segments the audio data flow into acoustically homogeneous segments, and the transcription system proper, which can be considered a large vocabulary continuous speech recognition (LVCSR) system with a number of possible acoustic model sets and language models. Transcription is carried out in a multi-pass framework where larger acoustic and language models are progressively introduced via recognition word graphs. Unsupervised speaker adaptation is carried out in the ultimate decoding pass.

4.1. *Automatic data partitioning*

While it is evidently possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straightforward solution. First, in addition to the transcription of what was said, other interesting information can be extracted, such as the division into speaker turns and the speakers' identities. Prior segmentation can avoid problems caused by acoustic discontinuities at speaker changes. By using acoustic models trained on particular acoustic conditions, overall performance can be significantly improved, particularly when cluster-based adaptation is performed. Finally, eliminating non-speech segments and dividing the data into shorter segments (which can still be several minutes long) reduces the computation time and simplifies decoding.

The data partitioning procedure, which is described more extensively in (Gauvain et al., 1998; Gauvain et al., 1999), aims at eliminating non-speech segments and at automatically segmenting the speech flow into acoustically homogeneous segments (wideband, telephone band, background noise, speaker...). Since there was no manually transcribed data available for German at the time this procedure was being refined, the German data have been segmented and labeled using the American English partitioner.

4.2. *Recognition system*

Acoustic model estimation

Gender-dependent acoustic models were built using MAP adaptation of speaker-independent seed models for wideband and telephone band speech. For computational reasons, a smaller set of acoustic models is used in the bigram pass to generate a word graph. The smaller sets contain about 1000 models (each with 3 states and 32 Gaussians per state) of position-independent, cross-word triphones covering about 40% of the triphone contexts. For trigram decoding larger sets of about 1500 position-independent, cross-word triphone models with a triphone coverage of around 50% are used.

These models have been trained for each phone set and pronunciation lexicon type (9 sets of about 1000 models for the bigram decoding pass and 9 sets of about 1500 models for the further decoding passes).

Language modeling

Language models are used to model regularities in natural language. The most popular methods, such as statistical *n*-gram models, attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of *n* words. A language model is obtained by interpolating multiple models trained on data sets with different linguistic properties. For example, commercially available broadcast news transcriptions, closed captions or subtitles, and newspaper and newswire texts, can be used to augment the transcriptions of the acoustic training data. Given a large text corpus it is relatively straightforward to construct *n*-gram language models. Most of the steps are relatively standard and make use of tools that count word and word sequence occurrences. The main considerations involve text normalization, the choice of the vocabulary and the definition of words, such as the treatment of compound words or acronyms, and the choice of the backoff strategy. In the experiments described here, bigram and trigram language models have been used. All language models used in the different steps were obtained by interpolation of backoff n-gram language models trained on different data sets.

Vocabulary selection

Over 300 M words of German text data (14 M sentences) were processed. Of these, about 2.6M words are distinct. However, many of the distinct lexical entries occur only once (54%). Table 5 shows the lexical coverage of the training texts as a function of the lexical size (the N most frequent words). Even with a lexicon containing 200K entries, almost 2.4% of the training words are unknown. This out-of-vocabulary (OOV) rate is much higher than observed in English and French, which is why we are looking into using morphological decomposition to increase the coverage for a fixed size lexicon (about 65k words). Table 5 shows the OOV rate on the German training data as a function of the lexical unit. The OOV rate using a recognition lexicon containing 65k words is 5.2%. Using a preliminary stemming procedure (including inflexion, suffix and prefix stripping, decompounding) to replace words by their stems, the OOV rate was reduced to 2.8%. The OOV rate was further reduced to 2.3% by ignoring case distinctions. For stemmed lexica no pronunciation dictionaries and language models were available yet. For the experiments reported here a case-sensitive 65k word recognition lexicon was used, without morphological decomposition.

Table 5. Lexical coverage achieved on the training text material using vocabularies of #words} most frequent words

#words	*coverage (%)*
10K	85.7
30K	91.7
65K	94.8
100K	96.1
200K	97.6

Word error metric

The commonly used metric for speech recognition performance is the “word error” rate, which is a measure of the average number of errors taking into account three error types with respect to a reference transcription: *substitutions* (one word is replaced by another word), *insertions* (a word is hypothesized that was not in the reference transcriptions) and *deletions* (a word is missed). The word error rate is defined as 100 × (#subs + #ins + #del) / #reference words, and is typically computed after a dynamic programming alignment of the reference and hypothesized transcriptions. Given this definition the word error rate can be more than 100%. Scoring is carried out using the Sclite scoring software from NIST. The scores reported here are prior to development of global mapping rules to correct for different commonly accepted orthographic forms (such as allowable alternative spellings for Genitive `-s` (`Papiers`, `Papieres`), compounded or uncompounded forms (`Kilometergeld`, `Kilometer Geld`)...

5. Experimental results

5.1. *Recognition results*

In Table 6, we report recognition results obtained with a trigram language model and unsupervised cluster-adapted acoustic models. All results are obtained using the same language models. Acoustic models depend on the pronunciation lexica and phone sets used. The number of parameters stay comparable across the different acoustic model sets.

Table 6. Word error rates on the 4 test programmes using different pronunciation lexica. For each programme the best result is put in boldface. Average results are given in the last line.

pron.lex.	**original**			**reduced**			**optional**		
show	**48**	**44**	**43**	**51**	**47**	**46**	**51**	**47**	**46**
news:									
arte_97:01:29-30	23.6	23.9	23.4	**23.2**	**23.2**	23.7	23.3	24.0	23.8
arte_97:01:20-23	20.5	19.2	20.2	19.6	**18.6**	19.6	19.7	20.0	19.8
documentaries:									
arte_98:09:28	23.5	23.2	23.0	23.3	23.9	**22.9**	24.2	23.9	24.1
arte_99:02:09	**28.4**	29.2	28.8	28.6	28.8	28.8	29.5	29.1	28.9
all shows	24.0	23.9	23.9	23.7	**23.6**	23.8	24.2	24.3	24.2

Various acoustic word modeling options were explored, either by using a larger or smaller set of phones or by the means of different or additional pronunciations. The word errors show only small variations in performance across the different configurations. Recognition results are slightly better when using the reduced pronunciation lexica.

5.2. *Discussion of errors*

Looking into the recognition errors in more detail, different sources may be distinguished which are related to the above-mentioned sources of lexical variation in German (and more thoroughly described in our companion paper in this workshop). Errors can be described using linguistic specificities of German or using more language-independent error classes.

Inflexions and derivations

Inflected forms of a given root form are likely to produce confusion errors. For articles and adjectives the `-em` ending (Dative sing.) is often replaced by the `-en` ending (Accusative sing., Dative plural) (examples of such confusions: `dem`, `einem`, `diesem`, `mittlerem`, `möglichem`, `unbestreitbarem`...). The Dative → Accusative confusion is about 3 times more frequent than the inverse Accusative → Dative substitution. The `-en` form is observed more often, hence better predicted by the language model. The `-em` form is often missing from the vocabulary and thus this type of confusion is often caused by an OOV problem. Another tendency is to replace longer forms by shorter forms (e.g. `sichere` by `sicher`, `vielversprechendsten` by `vielversprechenden`). This may be partially attributed to reduction phenomena, but also to insufficient lexical coverage (OOV problem).

Table 7. Error examples involving compounds. The **comment** indicates whether the **reference** word was missing in the vocabulary (OOV).

reference	**hypothesis**	**comment**
Juppé	Juppe	
Gasproduzenten	Gas Produzenten	OOV
Stundenwoche	Stunden Woche	
Parteienkonsenses	Parteien Konsenses	OOV
Bundeslandwirtschaftsministerium	Bundesland Wirtschaftsministerium	OOV
Präsidentenehepaar	Präsidenten Ehepaar	OOV
Weltwährungsfond	Welt Währungsfond	OOV
vorausgehen	voraus gehen	OOV
Verwaltungsfachleute	Verwaltungs Fachleute	OOV
Bilderwelten	Bilder Welten	OOV
Multimediataumel	Multimedia Taumel	OOV

Compounds

There are many examples of compounds being recognized as a sequence of separate items, mainly because the compound is missing, sometimes because it is too sparsely observed in the given context to be favorably predicted by the language model. Some of the errors are reported in Table 7. Errors mainly involve nouns. We can also analyse the errors using more language-independent error classes.

Short words

Short monosyllabic words are by far the most frequent words, mainly articles and prepositions (der, die, und, in, den, von, zu, mit, das, des, sich, auf, für...). But monosyllabic words can be found in all word classes: nouns (Zeit, Teil, Tag...) and proper names (Rom, Franz, Blair...), verbs (hat, ist), adjectives (rauh, eng...). Small words are easily inserted or omitted. For example, the conjunction und is frequently inserted in place of the negation prefix un- (unlaienhaft recognized as und Leidenschaft) or inflexions (word-final -n).

OOVs

Out of vocabulary words can be divided into two main categories: regular German words (with inflexions, derivations and compounds) or proper names, often of foreign origin. We have already discussed the problem of compounds. We can cite some typical examples of inflexions and derivations: Ausgelassenheit has been recognized as aus Gelassenheit, Vorsätzen as vor setzen, planzten as planzen, Erlöses as Erlös es..., Weinkeller as Wein Keller. Of course, not all of these OOVs are recognized as homophone word sequences (e.g. Politskandalen recognized as Polizei Sandalen, keimt der Verdacht as kam der Verdacht...), but often a large part of the overall meaning remains in the recognized word sequence. Proper names tend to introduce a large number of errors (especially if they are of foreign origin). Even if these errors are accounted for with the same weight as regular German word errors, the quality of the transcribed string is often strongly degraded without any link or resemblance to the reference (uttered) sequence. For example the reference sequence Anouk Aimée und Sandrine Kiberlin has been recognized as An dem E. und sonnt ging die Berner, the sequence die Weinberge des Clos Vougeot as die Weinberge des Globus so, the president Clinton as könnten. There certainly remains some phonemic similarity, but on the lexical level no obvious link remains between the reference and the recognized string. Hence, further automatic indexing may be much more affected by proper name OOVs than by compound OOVs.

Homophones and near-homophones

Some observed errors correspond to homophone confusions (e.g. `fielen` recognised as `vielen`, `Seen` as `sehen`) or to near-homophones: `Herden` recognised as `Erden`. Confusions easily occur between the vowel /a/ and the diphthong /ɑ^j/. (`Einspruch` recognized as `Anspruch`, `an` recognized as `ein`... Errors between inflected forms of a given root form also come into this category.

6. Conclusions

This paper gives an overview of the development of our automatic transcription systems for German and reports on experiments using different phone sets and pronunciation lexica for acoustic modeling. Slightly better results were achieved using the reduced lexicon as compared to the original or optional pronunciation lexica. Further experiments are planned using complex consonant cluster reductions in the pronunciation dictionaries.

Concerning the German transcription system in general, we are presently working on improving the acoustic and language models to lower the word error rate, which is significantly higher than in our American English system. This difference in word error rates can be attributed to several sources. First, there is a much higher lexical variation and variability in German than in English. Second, there is substantially less acoustic and textual data available for training the models. And thirdly, different types of data are being processed. The ARTE documentaries appear to be more challenging to transcribe than the news programmes.

7. References

Adda-Decker, M. & Lamel, L. (1999). Pronunciation variants across systems, languages and speaking style. *Speech Communication* **29**, 83-99.

Adda-Decker, M., Adda, G., Lamel, L.F. & Gauvain, J.-L. (1996). Developments in large vocabulary, continuous speech recognition of German. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96),* Atlanta.

Barras, C., Geoffrois, E., Wu, Z. & Liberman, M. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. *Proc. 1st Int. Conf. on Language Resources and Evaluation (LREC'98),* Granada, 1373-1376.

Duden 6 (1990). *Das Ausspracheworterbuch.* Mannheim: Duden Verlag.

Gauvain, J.-L., Lamel, L.F., Adda, G. & Jardino, M. (1999). Recent advances in transcribing television and radio broadcasts. *Proc. 6th Conf on Speech Comm. and Techn. (Eurospeech'99),* Budapest.

Gauvain, J.-L., Lamel, L.F. & Adda, G. (1998). The LIMSI 1997 Hub-4E transcription system. *Proc. DARPA Broadcast News Transcription & Understanding Workshop,* Landsdowne, 75-79.

Lamel, L.F., Adda-Decker, M. & Gauvain, J.-L. (1995). Issues in large vocabulary, multilingual speech recognition. *Proc. 4th Conf on Speech Comm. and Techn. (Eurospeech'95),* Madrid.

Rolduc (1998). Workshop on Modeling Pronunciation Variation for ASR. *ESCA-ETRW,* Rolduc.

SpeechCom (1999). Special Issue on Pronunciation Variation Modeling. *Speech Communication* **29**.

Young, S., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J.-L., Kershaw, D., Lamel,, L., Leeuwen, D., Pye, D., Robinson, A., Steeneken, H. & Woodland, P. (1997). Multilingual large vocabulary speech recognition: the European SQALE project. *Computer Speech and Language* **11**(1), 73-89.