Underspecified recognition

Aditi Lahiri and Henning Reetz

Abstract

The FUL (Featurally Underspecified Lexicon) model assumes phonological representations of morphemes with hierarchically structured features, not all of which are specified. Such underspecified representations are assumed for the mental lexicon as well as for the computerised lexicon employed for automatic speech recognition. In FUL, a segment is lexically represented by sufficient features to separate it from any other segments in the phonology of a particular language. In speech production, adjacent features 'fill in' underspecified slots, thereby accounting for assimilations. In speech perception, incoming speech sounds are compared online to these sets of features with a ternary logic of match, mismatch, and no-mismatch. Features that are present in the acoustic signal do not mismatch with the underspecified (i.e. 'empty') slots in the lexicon. In such an approach, speech perception can deal with different kinds of within- and across-speaker variation found in normal speech, without listing every variant in the lexicon. Along with diachronic data and the results of psycholinguistic experiments, the computational performance of our automatic speech recognition system successfully demonstrates the adequacy of this model.

1. Introduction

The speech signal of the same phonetic segment varies across dialects and speakers, within speakers between segmental and prosodic contexts, and even for the same speaker and context with repetition, speaking rate, emotional state, microphone and line condition, etc.* Ambiguities in the signal, whether they come from random noise or whether they are linguistic in nature, like cliticisations of words, or assimilations, partial or otherwise, are the norm rather than the exception in natural language. Human listeners, however, appear not to be too concerned by adverse acoustic conditions and indeed, handle "variations" in the signal with aplomb. Any theory of lexical phonological representation and recognition must be able to account for productive phonological processes such as assimilations, particularly across word boundaries. Explicitly or implicitly, all such theories assume that at the level of the lexical entry there is a single abstract representation, so that not every phonological surface variant form is listed.[1] This leaves unanswered, however, the question of precisely how the system does recognise the different phonetic variants of a word when the relationship between these realisations and the lexical entry is not straightforward.

We will consider here the linguistic, psychological, and computational adequacy of our approach to this question. The approach we advocate assumes a featurally underspecified lexicon, extraction of features from the acoustic signal, and a ternary matching condition which matches the output features to the lexically specified features. The predictions of our model – FUL (Featurally Underspecified Lexicon) – are evaluated on the basis of language comprehension experiments, evidence from language change, and its computational performance in an automatic speech recognition system. The crucial assumptions of FUL are given below.

(1) Underspecified recognition: the FUL model

a. The phonological representation is abstract and underspecified. The feature representation for each segment is constrained by universal properties and language specific requirements.

b. Each morpheme has a unique representation. No phonological variants, morphophonological or postlexical, are stored.

c. The perception system analyses the signal for rough acoustic features which are transformed into phonological features. There is no conversion into segments or syllables and there is no further intermediate representation.

d. The phonological features are mapped directly on to the lexical representation. A three-way matching condition (*match, mismatch, no-mismatch*) determines the choice of candidates activated. Along with the phonological information, morphological, syntactic and semantic information is made available.

Each point is discussed briefly in turn. (1a) A segment is represented with a root node and its relevant features, similar to that presented in Lahiri & Evers (1991), Lahiri (2000) and Ghini (2001). The most salient aspects of this representation are that (i) features are privative or monovalent, (ii) vowels and consonants share the same place features, and (iii) the place features split into two nodes: the articulator node consisting of the places of articulation, and the height features under the tongue height or aperture node. Hence [HIGH] and [LOW] (height features) are independent of the places of articulation [LABIAL], [CORONAL] and [DORSAL].

Not all features are represented in the lexicon. The specification of features depends both on universal and language specific grounds. For instance, the FUL system has the feature [ABRUPT] in its inventory, but it is not specified in the lexical representation for German morphemes. Neither is the feature [CORONAL] specified. The assumption is that features like [ABRUPT] and [CORONAL] are left unspecified unless the phonological system of the language requires it. In our model, underspecification is context-free.

Underspecification and underspecified representations have been the source of considerable dissension.[2] A recent critique, reexamining the pros and cons of the issues and providing further evidence in favour of underspecification, is given in Ghini (2001). Ghini shows that a complex pattern of vowel alternation in the dialect of Miogliola supports not only the underspecification of vowel features, but also that two superficially similar dental nasals are underlyingly different – one specified for [CORONAL] which always surfaces as [n], and the other unspecified for place which surfaces as [ŋ], [ɲ] and [n] under prosodically defined conditions. These facts are discussed in more detail with reference to language change in § 3.

The next assertion (1b), that no phonological variant is stored, is also linked to the notion of an underspecified representation. No postlexical variants are stored and if morpheme alternants are phonologically related the assumption is that only a single underlying representation is available. From the signal to the phonological representation, the perceptual system extracts rough acoustic characteristics which are converted into phonological features (1c). All features are extracted independent of whether they are specified or unspecified in lexical representations. The features extracted from the speech signal are then compared to those stored in the lexicon. There is no conversion from features into segments; in fact there is no intermediate representation of segments, syllables or any other phonological unit. The mapping from the features to the representation entails a ternary system of matching (1d): *match, no-mismatch* and *mismatch.* The *match* condition can only occur if both signal and lexicon have the same features. This condition is used for the scoring of word candidates and includes a correction formula to account for different sized feature sets. The *mismatch* occurs if signal and lexicon have contradicting features. A mismatch excludes a word from the list of possible word candidates. The mismatching relationship can be bidirectional. For instance, [HIGH] and [LOW] mismatch, independent of which is extracted from the signal and which is stored in the lexicon. A mismatch can occur also in one direction due to underspecification. For example, any one of the place features [LABIAL], [DORSAL], or [CORONAL] can be extracted from the signal but only [LABIAL] and [DORSAL] are stored in the lexicon. If the feature [CORONAL] is extracted from the signal then it mismatches with the features [LABIAL] and [DORSAL]. The other way round, the signal feature [LABIAL] *does not* mismatch with an underspecified coronal sound.

A *no-mismatch* occurs (i) if no feature is extracted from the signal that is stored in the lexicon, or (ii) if a feature is extracted from the signal that is not stored in the lexicon, and (iii) by definition (e.g. [CORONAL] does not mismatch with [HIGH]). Case (i), where no feature is extracted from the signal but features are available in the lexicon, does not lead to a rejection of candidates. The signal simply does not contradict a candidate; only the candidate does not increase its matching score (see below). Case (ii) is exactly the case for the lexical feature [CORONAL]: coronality is not stored in the lexicon. If a place feature like [LABIAL] or [DORSAL] is extracted from the signal, it does not mismatch with a coronal sound in the lexicon.

(2) Examples of *match, mismatch* and *no-mismatch*

Signal	Matching	Lexicon
[HIGH]	*mismatch*	[LOW]
[CORONAL]	*mismatch*	[DORSAL]
[DORSAL]	*no-mismatch*	[UNSPECIFIED PLACE]
[DORSAL]	*match*	[DORSAL]

All word candidates that agree (match or do not mismatch) with the initial feature set are activated, together with their phonological, morphological, syntactic, and other information. Matching features increase the activation level for potential word candidates, non-mismatching features do not exclude candidates and only mismatching features lead to the rejection of word candidates. The level of activation is measured on the basis of the number of matching features with respect to those specified in the lexicon and the number of features extracted from the signal (Reetz 1998). Each candidate receives a score on the basis of the formula given in (3):

(3) Scoring formula

$$\text{SCORE} = \frac{(\text{NR. OF MATCHING FEATURES})^2}{(\text{NR. OF FEATURES FROM SIGNAL}) \times (\text{NR. OF FEATURES IN THE LEXICON})}$$

To illustrate the scoring method, the features extracted from the signal for the first vowel in the intended German word *müde* [myːdə] 'tired' would optimally be [HIGH, CORONAL, LABIAL]. The features in the lexicon are [HIGH, LABIAL]. Given these features, the scores of other front vowels would be as follows:

(4) Scores for [y]

	Lexical features	Input features of [y]	Score
[y]	[HIGH, LABIAL]	[HIGH, CORONAL, LABIAL]	$2^2/(3 \times 2) = 0.66$
[i]	[HIGH]		$1^2/(3 \times 1) = 0.33$
[Y]	[HIGH, LABIAL, RTR]		$2^2/(3 \times 3) = 0.44$
[ɪ]	[HIGH, RTR]		$1^2/(3 \times 2) = 0.16$

According to these scores, a word like *Mücke* [mʏkə] 'mosquito' would be a higher scoring candidate than *Miete* [miːtə] 'rent' for the initial sequence [myː] of *müde*. If for some reason the feature [LABIAL] was not present in the signal, or could not be extracted by the listener, the FUL system predicts that [i] would have the highest score but [y] would still be available. None of the low vowels, however, would be considered if [HIGH] was extracted since it would mismatch with lexically specified [LOW]. We will discuss this in more detail in § 4.

In the next three sections we briefly go through some data in support of the FUL model from language comprehension experiments, language change, and finally from the speech recognition system that we are developing.

2. Underspecification in language comprehension

In this section, we focus on the adequacy of the assumptions in FUL for language comprehension. To this end we will summarise some experiments incorporating the concept of underspecification for lexical access.

As we mentioned above, assimilation can lead to surface variants. Assimilation of a coronal sound (e.g. /n/) to a following labial place of articulation (like /b/ in "Where could Mr. Bean be?") often results in the production of a labial (i.e. *Bea[m] be*). The reverse is not usually true, that is, a labial sound does *not* assimilate to a coronal place of articulation (i.e., *la[m]e duck* does *not* become *la[n]e duck*).[3] Simple articulatory mechanics cannot account for such behaviour because an articulatory assimilation would operate in both directions. An explanation can be given by assuming that coronal sounds are underspecified for place, whereas labials and dorsals are not: the labial place of articulation spreads to the preceding coronal sound (if the language has regressive assimilation) because that sound is not specified for place. On the other hand, the specification of a labial place prevents the place features of an adjacent sound from overriding this information. Consequently, coronal sounds can become labial (or dorsal), but labial (or dorsal) sounds do not change their place of articulation.

This explanation is straightforward for speech production, but not so in speech perception. How can a realisation of *gree[m]* in a labial context (like *bag*) or *gree[ŋ]* in a dorsal context (like *grass*) lead to the access of the word *green* in the lexicon? The utterances *gree[m]* and *gree[ŋ]* are nonwords in English. And, at the same time, how should a mechanism be constructed to disallow the activation of the word *bean* if the acoustic input is *bea[m]*, even if *bean* is a word of the language? Human listeners handle these asymmetries (and many other assimilatory effects) within and across words without noticing it, as experimental evidence indicates (Lahiri & Marslen-Wilson, 1991; Gaskell & Marslen-Wilson, 1996; Gow, 2001; Lahiri & van Coillie, to appear). The solution to these seemingly contradictory requirements can be obtained in the FUL system by assuming an underspecified representation in the lexicon, where certain features (like the place feature [CORONAL]) are *not* stored in the lexicon (in speech production, segments with unspecified place are generated with the feature [CORONAL] by default) and by postulating the ternary matching logic in the signal-to-lexicon mapping.

Assuming that phonological lexical representations of words consist of underspecified featural representations, Lahiri & Marslen-Wilson (1991, 1992) argue that the mapping process from the signal to the lexicon crucially depends on the absence and presence of features in the representations of words in the mental lexicon. They contrasted vowel nasality in Bengali and English, where Bengali has underlying nasal vowels as well as contextual nasalisation ([bʰɑ̃r] 'clay bowl', /bʰɑn/ > [bʰɑ̃n] 'pretence'). On the other hand, any nasality on a vowel in English comes from a neighbouring nasal consonant. They argued that only underlying contrastive nasal vowels in Bengali are specified for nasality; for other vowels, no nasality is specified. Results show that indeed, the listener always interprets nasality on a vowel as being contrastively nasal even if the stimulus segment contained a vowel which was contextually nasalised. More strikingly, oral vowels in oral contexts, for both English and Bengali (English *bad*, Bengali [bʰɑr] 'weight'), are interpreted by listeners as having either a nasal or an oral context, depending on the distribution of the words in the language. That is, the vowel [æ] in *bad* was equally likely to be a interpreted as being part of *bad* or *ban*, showing that in both languages the oral vowels were represented as unspecified for nasality in spite of the fact that there may be surface phonetic nasalisation present in production of CVN words.

Thus, along with underspecification, the three-way matching – match, no-mismatch and mismatch – gives the asymmetry between coronals on the one hand and labials and dorsals on the other. If [CORONAL] is extracted from the signal, then it mismatches with [LABIAL] (i.e. [n] mismatches with underlying /m/). It does not find a perfect match since /n/ is not specified for [CORONAL], but it does not mismatch either – hence a no-mismatch situation occurs. If [LABIAL] is retrieved from the signal, it matches perfectly with underlying /m/, but it also does not mismatch with /n/. This is not the best match, but it is a no-mismatch. Examples of the three way distinction are given in (5).

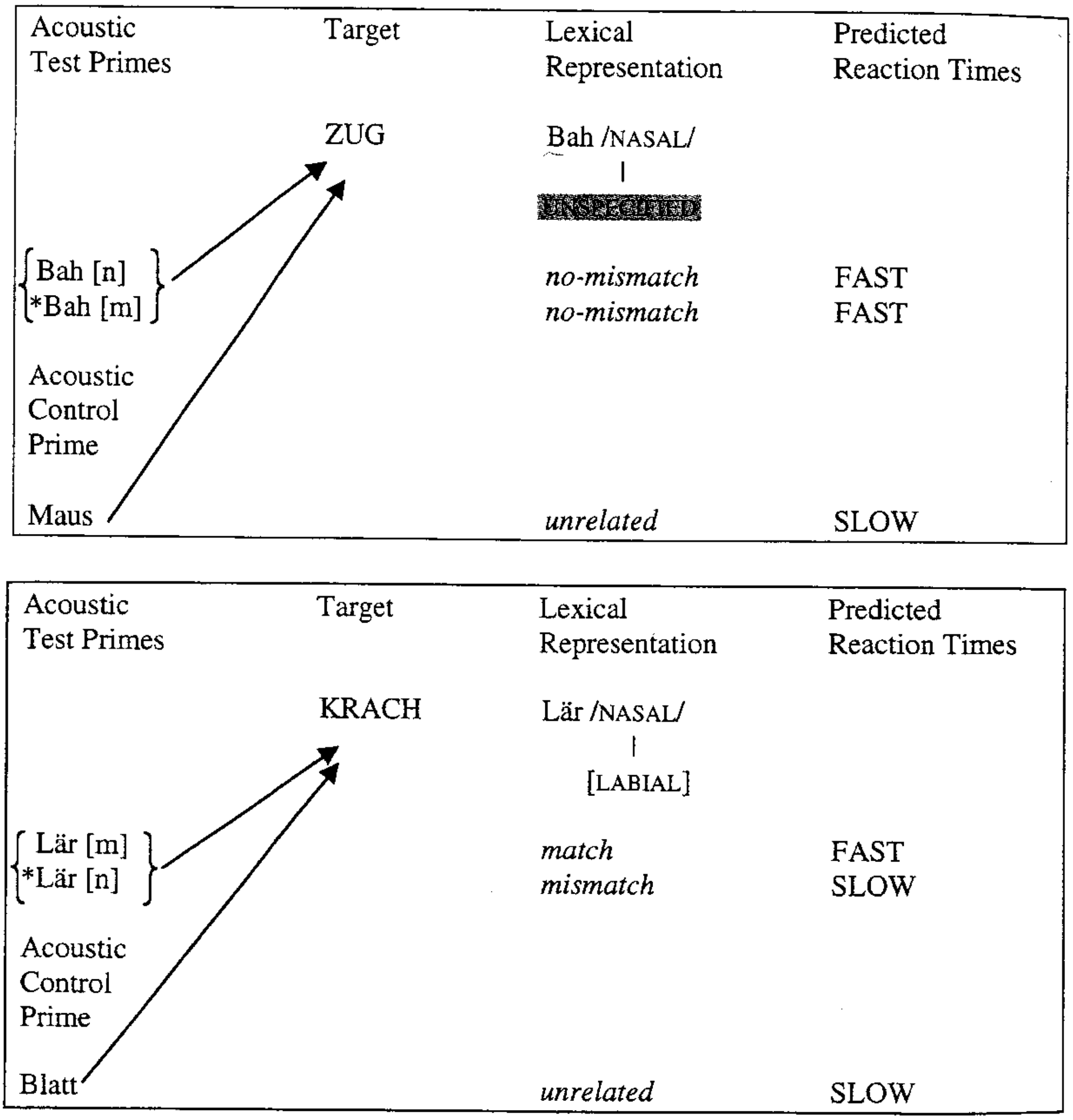
(5) Matching from signal to lexicon

Signal	Matching	Lexicon
[HIGH]	*match*	[HIGH]
[STRIDENT]	*mismatch*	[NASAL]
[LATERAL]	*mismatch*	[NASAL]
[CORONAL]	*mismatch*	[LABIAL]
[CORONAL]	*no-mismatch*	[UNSPECIFIED PLACE]
[LABIAL]	*no-mismatch*	[UNSPECIFIED PLACE]

The system can handle within and across word assimilations and can deal with a certain number of dialectal variants. The asymmetry in assimilation is explained by the fact that since the place feature [CORONAL] is unspecified, the feature [LABIAL] detected in the signal for *gree[m] book* does not mismatch with the lexical representation *green*. However, the feature [DORSAL] detected in the signal for *ho[ŋ]* does mismatch with the feature [LABIAL] in the lexical representation of *home,* and so *home* is rejected. Coronals get a lower score than the labials (or dorsals), which obtain a match, but coronals are not excluded. They remain active as assimilatory variants.

Lahiri & van Coillie (to appear) provide further evidence for the underspecification of [CORONAL], and the efficacy of the three-way matching. We will briefly discuss two experiments. In both experiments, a crossmodal lexical decision task with semantic priming was used. The listeners were auditorily presented with a real word like *Bahn* 'railway' or *Lärm* 'noise' in isolation. At the offset of the acoustic stimulus, the subjects saw a semantically related word like *Zug* 'train' or *Krach* 'bang, racket' and they had to decide whether it was a word or not. Since it is well established in the psycholinguistic literature that semantically related words prime, the expectation was that the subjects would be faster in reacting to *Zug* after they heard *Bahn* as compared to an unrelated word *Maus* 'mouse'. Similarly, *Krach* should be recognised faster after *Lärm* rather than after *Blatt* 'leaf'. The question of course is whether nonword variants of the real word primes would have any effect. That is, would related acoustic variants **Bahm* or **Lärn* prime *Zug* and *Krach* respectively? The first experiment presented here examined word final nasals. The experimental design and predictions are illustrated below.

(6) Recognition of word final nasals: predictions



The claim is that although the acoustic prime **Bahm* is a nonword it does not mismatch with the lexical representation of *Bahn* and therefore successfully activates *Zug*. The signal has the feature [LABIAL] but the lexical representation has no place specified in the lexicon and hence it is not rejected. This is not the case with an underlying labial. When the feature [CORONAL] is extracted from the nonword **Lärn*, it mismatches with the lexically represented [LABIAL] of the real word *Lärm* and hence its semantic associate is not activated. Apart from the nasal, we used additional consonantal variations as primes which deviate from the lexical representation with respect to other features as well. The matching expectations are given in (7).

(7) Other consonantal variants of word final nasals

a. Variants of final /-n/

Acoustic variant	Acoustic features	Lexical Representation of /n/	Matching
Bah[l]	[LATERAL]	[NASAL]	*mismatch*
	[CORONAL]	[UNSPECIFIED PLACE]	*no-mismatch*
Bah[p]	[ABRUPT]	[NASAL]	*mismatch*
	[LABIAL]	[UNSPECIFIED PLACE]	*no-mismatch*
Bah[s]	[STRIDENT]	[NASAL]	*mismatch*
	[CORONAL]	[UNSPECIFIED PLACE]	*no-mismatch*

b.Variants of final /-m/

Acoustic variant	Acoustic features	Lexical Representation of /m/	Matching
Lär[w]	[CONTINUANT]	[NASAL]	*mismatch*
	[LABIAL]	[LABIAL]	*match*
Lär[p]	[ABRUPT]	[NASAL]	*mismatch*
	[LABIAL]	[LABIAL]	*match*
Lär[s]	[STRIDENT]	[NASAL]	*mismatch*
	[CORONAL]	[LABIAL]	*mismatch*

Note that features like [ABRUPT] and [CORONAL] are not specified in the German lexicon. It does not mean that this is always so.

Depending on the grammar of a particular language, [CORONAL] for instance can be specified for nasals and not for stops, as laid out in the next section.

As indicated in (6) and (7), for each word, 4 nonword primes were created by changing the final consonant.[4] Thus for the word *Bahn*, the nonwords were *Bah[m], Bah[l], Bah[s], Bah[p].* In addition, there was a control prime for each target, where the control was unrelated to the real word prime (e.g. *Maus-Zug*). Thus, in all, for one target there were six different primes. However, each subject was faced with the target only once. Reaction time measurement started at the offset of the auditory prime when the visual target was presented. The results of the experiment are given in Figures 1 and 2. In Figure 1, we see the reaction times when the prime was a variant of an underlying /-n/ unspecified for place. In comparison to the control, there is a significant priming effect for the real word *Bahn* as well as the variant **Bahm* where the final consonant did not mismatch.[5] In all other instances there was no priming effect. Recall that under our assumptions, there is no difference in matching between the surface *Bah[n]* and *Bah[m]* when compared to the real word *Bahn* (see 6). Since the nasal is unspecified for place, both variants with [n] and [m] are no-mismatches. Hence we did not expect any difference between the two conditions and our results bear this out – they do not differ significantly.

In Figure 2, we see the results of the variants of the words with final /-m/. Here, the only word which has a significant priming effect is the real word *Lärm.* In contrast to the word final [n] as in *Bahn,* where its labial variant **Bahm* also caused priming of *Zug*, the variant **Lärn* did not prime the semantically related word *Krach* of the real word *Lärm.* Moreover, there was a significant difference between *Lärm* and **Lär[n].* The clear difference in the results supports our expectations regarding underspecified features and the match/no-mismatch asymmetry.

The next experiment examined word medial nasals. In the case of word final nasals there is a possibility of assimilation, but word medial nasals, particularly intervocalic nasals, remain untouched. Hence, one could argue that even if underspecification was a reasonable choice for word final nasals, such an option would be

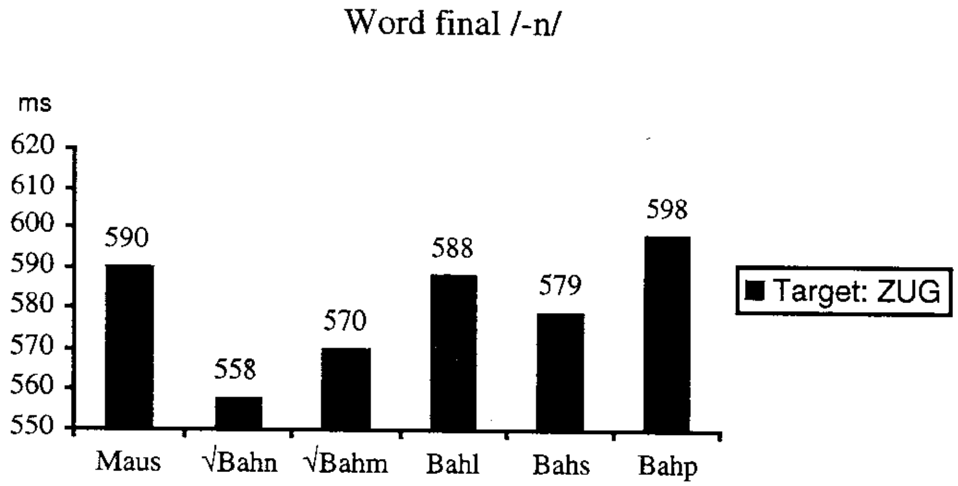


Figure 1: Mean reactions times to a semantically related target for a word ending in /-n/ and its variants (each class consists of 24 words represented here with one example). Significant priming effects with respect to the control are indicated by √.

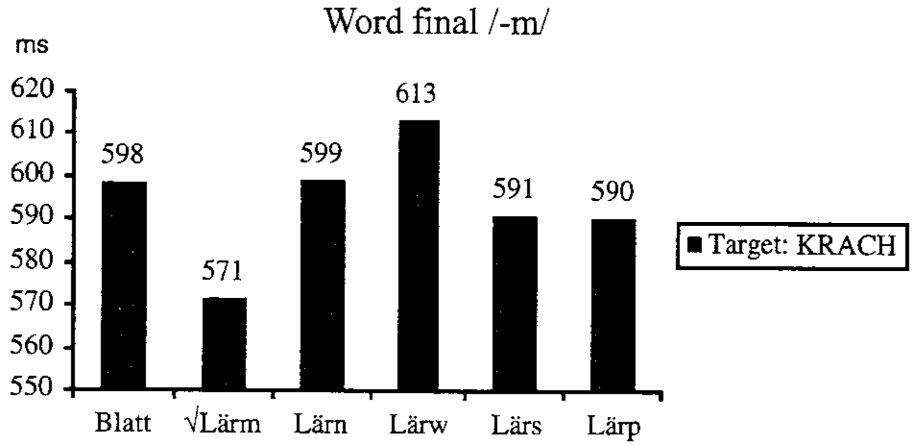


Figure 2: Mean reaction times to a semantically related target for a word ending in /-m/ and its variants (each class consists of 12 words represented here with one example). Significant priming effects with respect to the control are indicated by √.

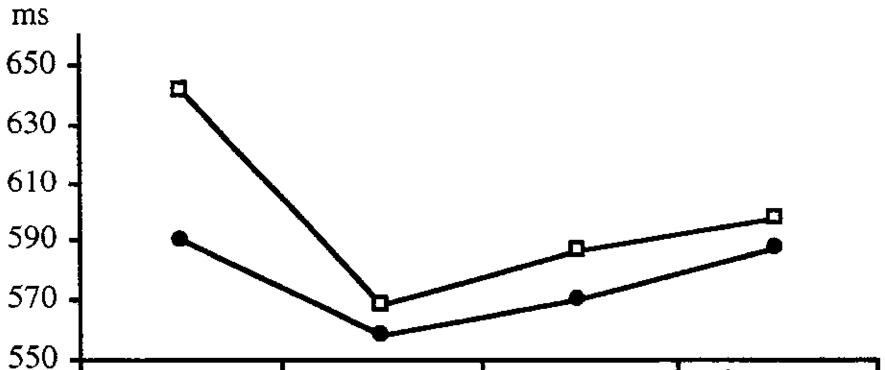
unnecessary in medial position since there is no possibility of alternation due to assimilation. Under the FUL model, underspecification is not determined by the position in a word. Coronal consonants are unspecified for place no matter what position in a word they occur in. Our prediction is therefore, that the same asymmetry would hold for word medial nasals just as like the word final ones.

In the second experiment, assuming that medial coronal consonants in German are all unspecified for place, both obstruents as well as nasals were examined. For the sake of comparison, only the nasal data are discussed. The task was the same, and the real word primes with medial nasals were converted into two different types of nonwords: a nasal with a different place of articulation and a non-nasal consonant (*Düne* 'dune', **Dü[m]e*, **Dü[l]e*; *Schramme* 'a scratch', **Schra[n]e*, **Schra[v]e*). The targets, as before, were semantically related, and were presented visually at the offset of the prime.[6]

The results of the second experiment confirms our earlier findings. *Düne* 'dune' primes its semantically related word *Sand* 'sand' just as well as nonwords made up with a non-mismatching [LABIAL] nasal (like **Dü[m]e*). In contrast, although *Schramme* 'a scratch' primes its semantically related word *Kratzer*, (also 'a scratch'), the nonword with a coronal (**Schra[n]e*) does not. The asymmetry again shows that when the feature [LABIAL] is extracted from the signal, it does not mismatch with the underlying unspecified [CORONAL], but an extracted [CORONAL] does mismatch with a [LABIAL]. There was no priming in the other nonword conditions. In the next two figures we compare the results for the word medial and word final nasals in the same graph.

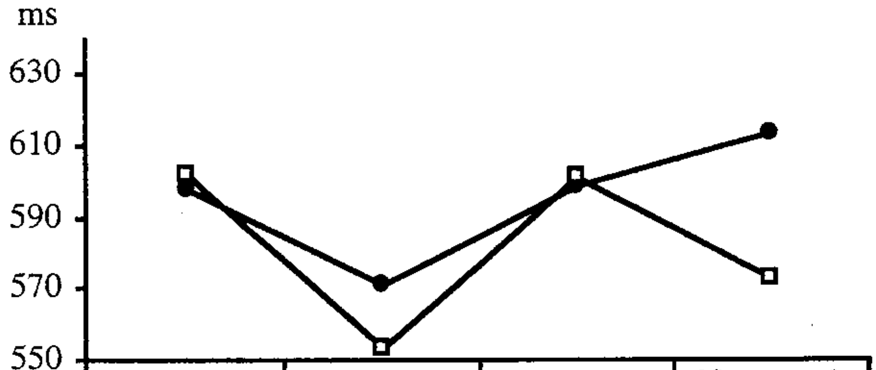
The pattern of results of the medial and final coronal nasals are very similar, as shown in Figure 3. Both the real words and their acoustic variants with [m] show significant priming effects with respect to the control. However, there is no priming with the other nonword primes as compared to the control words. This is not the same for the labials, as we can see in Figure 4.
Here, the real words with [m] are significantly faster than the control, but neither the nasal variants nor the non-nasal variants are significantly faster. The nonwords with coronal nasals mismatched and hence were no different from the unrelated controls, and they were also significantly slower than the real words with [m].

Thus, language comprehension experiments suggest that the predictions made by the FUL model combining underspecification with a three-way matching – perfect match, no-mismatch and mismatch – are borne out. The experimental results support the predicted asymmetry between coronals on the one hand, and labials and dorsals on the other.[7]



	Control	√ Word [n]	√Variant [m]	Non-nasal variants
—●—/-n/	590	558	570	588
—□—/-n-/	642	568	587	598

Figure 3: Comparing word final /-n/ and word medial /-n-/. Mean reaction times to the real word primes with the coronal nasal, the primes with the non-mismatching nasal variant [m], and the primes with mismatching non-nasal variants. Significant priming is indicated by √.



	Control	√ Word [m]	Variant [n]	Non-nasal variants
—●—/-m/	598	571	599	613
—□—/-m-/	602	553	601	573

Figure 4: Comparing word final /-m/ and word medial /-m-/. Mean reaction times to the real word primes with the labial nasal, the primes with the mismatching nasal variant [n], and the primes with mismatching non-nasal variants. Significant priming is indicated by √.

3. Underspecification in language change

In general, the proponents and opponents of underspecification have leant on synchronic alternations to support their point. As we mentioned earlier, Ghini (2001) shows that underspecification of vowel and consonantal features is crucial for the understanding of the complex interaction of prosodic and segmental phenomena in Miogliola, an Italian dialect of Liguria. In particular, two superficially similar dental nasals are different in their underlying representation in terms of place specification. One of them is specified for the feature [CORONAL] and always surfaces as [n], while the other is unspecified for place and surfaces as a palatal, dental or velar, depending on context. If underspecification is part of the mental lexicon, as we claim, then it should play a role in language change since part of change is in fact building representations by a new generation. In Ghini (to appear), we find additional support for this representation from language change. The crucial facts are summarised below.

Many Italian dialects like Miogliola have lost the quantity distinction in both obstruents and nasals. Thus all Latin geminates are single consonants in these dialects. For obstruents, however, a further process of spirantisation along with voicing has helped to maintain the distinction between original Latin single consonants and geminates in Miogliola. If we compare Latin, Standard Italian and Miogliola, we find that Latin [p] remained in Italian, but became [v] in Miogliola. In general, Latin single voiceless stops and fricatives became voiced fricatives in Miogliola ([p, f] > [v]; Latin *lupus* 'wolf', Miogliola [lūv]). Geminate obstruents, however, simply degeminated in Miogliola and the stop has not undergone spirantisation. That is, Latin labial stops and fricatives [pː, fː] became simple [p, f] (Latin *cippus* 'pillar', Miogliola [ʦæp]). This is shown in (8a). Thus, the original geminate/single consonant contrast of Latin is now maintained as a stop/voiced fricative contrast in Miogliola.[8]

In sonorants, there was no possibility of voicing or spirantisation, and hence, after degemination, there was a general neutralisation of the length contrast: Latin [mː, m] > Miogliola [m]. As we see in the examples in (8b), there is a single labial nasal in Miogliola now and the original geminate/nongeminate [mː]/[m] contrast of Latin has been neutralised. In standard Italian, the original length contrast is still maintained. The spirantisation and voicing processes could play no role in the case of the labial nasal.

(8) Loss of geminate/nongeminate contrast from Latin to Miogliola

Classical Latin		Italian	Miogliola	
a. Latin length distinction changed to segmental distinction for obstruents.				
ci**pp**us	'pillar'	ʧe**pp**o	tsæ**p**	tsæ**p** + ɪ (PL)
lu**p**us	'wolf'	lū**p**o	lū**v**	lū**v** + ɪ (PL)
b. Sonorants did not lenite – distinction lost with labial nasals				
su**mm**us	'utmost'	so**mm**o	su**m**	su**m** + ɪ (PL)
fū**m**us	'smoke'	fū**m**o	fy**m**	fy**m** + ɪ (PL)

However, given the assumption of underspecification, the coronal nasal had the possibility of a dual pattern of change. And this is what happened: the original quantity distinction was transformed to a place distinction. The following examples illustrate this.

(9) QUANTITY distinction to place distinction for coronal nasals

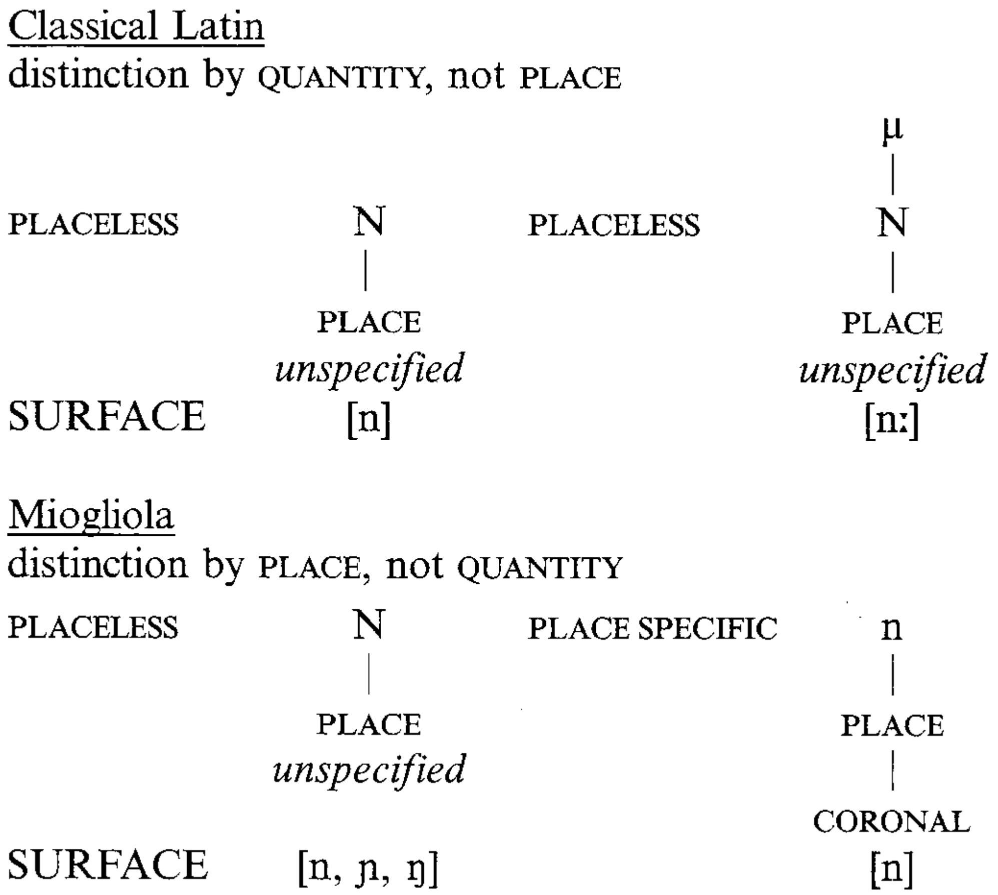
Classical Latin		Italian	Miogliola				
pa**nn**us	'cloth'	pa**nn**o	pā**n**		pā**n** + ɪ		(PL)
ca**n**is	'dog'	kā**n**e	ka**ŋ**		ka**ɲ** + ɪ		(PL)
alu**mn**us	'alumni'	ȧlu**nn**o	alý**n**	MASC	alý**n**a	FEM	(SG)
			alý**n**ɪ	MASC	alý**n**ɛ	FEM	(PL)
ū**n**us	'one'	u**n**o	ø**ŋ**	MASC	œ̃́**n**a	FEM	(SG)
			ý**ɲ**ɪ	MASC	œ̃́**n**ɛ	FEM	(PL)

LATIN		MIOGLIOLA
geminate [**nː**]	>	[**n**]
single [**n**]	>	[**ɲ**] in onset followed by [ɪ], [**ŋ**] in coda, [**n**] elsewhere

Both the single /n/ as well as the geminate /nː/ in Latin were unspecified for place. In Miogliola, the original Latin single coronal nasal /n/ remained unspecified for place – that is, there was no change. The geminate /nː/ degeminated, but became specified for place. As a result, the synchronic grammar of Miogliola shows surface neutralisations from two underlyingly different specifications. Latin *pan-*

nus became Miogliola [pɑ̄n] SG., [pɑ̄n + ɪ] PLURAL, where Miogliola has lost the geminate/nongeminate contrast. But the [n] in Latin *ūnus* has several surface variants in Miogliola: [øŋ], [œ́na], [ýɲɪ] and [œ́nɛ]. The quality of the nasal depends on whether it is in the coda and on the quality of the following vocalic suffix. Compare now the original Latin *alumnus,* where the /-mn-/ sequence became a geminate at a later point. Here, all four gender and number contrasts are also present as in 'one', but the consonant is always a dental [n]. Note the differences in the masculine plural: [ýɲɪ] (from unspecified /N/) and [alýnɪ] (from specified /n/). The quality of the vowel has no effect on the original geminate /nː/. The change in the representation from Latin geminate coronals to Miogliola is illustrated in (10).

(10) Change of coronal nasals from Latin to Miogliola due to underspecification



As can be seen, the placeless /N/ can take place features according to segmental and prosodic contexts and surface as [n, ɲ, ŋ], while the coronal nasal specified for place, surfaces always as a dental [n].

Thus, synchronically, Miogliola has two coronal nasals, only one of which is specified for place. Like many other languages, underspecification is used contrastively. The history of these nasals show that the source of the place-unspecified nasal is the original nongeminate coronal, which maintained its underspecification and has several surface variants depending on prosodic contexts. It is the original geminate coronal nasal which became a single consonant and acquired place specification. This consonant has no surface variants. What is interesting is that it was possible for the language learner to take advantage of the underspecified place representation to maintain the original geminate/nongeminate contrast. This was not possible for the labial nasals which were already specified for place.

Thus, under our view, underspecified phonological representations, being a part of the mental representation, play a role for both processing and change. Some of the notable parallel aspects are summarised in the following table.

(11) Processing and Change with respect to underspecified phonological representations

PROCESSING	CHANGE
a) Segments can vary according to context, leading to loss of contrast; however, there is asymmetry in the variants.	a) Sound change can lead to loss of contrasts and restructuring; however, there is occasional asymmetry in restructuring.
b) Asymmetry in representation leads to asymmetry in recognition.	b) Asymmetry in representation is reflected in phonological change.
c) Underspecified representations lend themselves to a three-way matching with features from the signal, allowing for the recognition of neutralised segments.	c) Underspecified representations can be exploited by language learners to maintain contrasts which would have otherwise been neutralised.

4. The FUL model of speech recognition

In an attempt to model a system with an underspecified lexicon and a three-way matching described above, we have developed an automatic speech recognition system which runs on these lines (Reetz, 1998, 1999; Lahiri, 1999). The central goal behind this enterprise is to test the actual viability of a feature based extraction system in combination with an underspecified lexicon and a ternary matching condition. Experimental results in language comprehension allowed us to believe that the human system does not use a fully specified phonological representation and that there is an asymmetry in the matching from the signal to the lexicon. Evidence from language change also suggests that the asymmetry in place representation can lead to an asymmetry in the restructuring of forms and to the establishment of an altered pattern of contrasts. Both pieces of evidence are real but do not provide us with a handy means of testing the predictions. We therefore took on the task of building a model based on our premises, with the addition of an acoustic front-end which could handle the online extraction of features.

Given the variation in the speech signal, it is not surprising that automatic speech recognition using simple spectral template matching has problems. Any variation of the signal leads to variation of the spectra that are compared to the stored templates. Klatt (1989) provides a comprehensive review of models which endeavour to solve the variation problem by storing all spectral information in the lexicon. The more popular approach to resolving such variation is a statistical one. Statistical approaches like Hidden Markov Models based on large training sets have led to acceptable results, but are still speaker and transmission-line dependent. Moreover, the success of the HMMs depend more on probabilities of longer strings of data (including word sequence probabilities) rather than on a front-end phonetic analysis. The system presented here operates on completely different principles, both with respect to the front-end as well as the lexicon. No spectral templates are computed from the speech signal to access the lexicon. Neither is the signal analysed in great detail for acoustic evidence of individual segments and their boundaries. The principal aspects of the FUL model are the following:

(12) Characteristics of the FUL speech recognition system

a. The system is based on the phonological representation of words in the lexicon.

b. Each word has a unique representation in spite of the large variation. The phonological representation is feature-based and assumes underspecification.

c. The speech signal is converted into distinctive phonological features. The conversion operates speaker-independently and without prior training.

d. Once the features are extracted the system never re-evaluates the acoustic signal, i.e. there is no close phonetic investigation of the signal to verify or falsify word hypotheses.

e. Features extracted from the signal are matched with those stored in the lexicon using a ternary system of matching, non-mismatching, and mismatching features. All word candidates that match with the initial feature set are activated, together with their phonological, morphological, syntactic, and other information.

f. The word candidates are expanded to include word hypotheses, even without complete acoustic evidence, which are then available for the phonological and syntactic parsing that uses additional prosodic and other information and operates in parallel with the acoustic front-end.

The lexical representations are similar to what we have seen before. Each morpheme is represented with root nodes linked to minimal feature specifications. For the sake of space, in the examples in (13) the features are listed in a linear string for each segment. In all, FUL requires twelve phonological features.[9] Quantity is represented in terms of moras and not by features.

(13) Lexical feature specifications for German

BAHN	/baːn/	
railway	/b/	[CONS] [LAB] [VOICE]
	/aː/	[LOW] [DORSAL]
	/n/	[NASAL]

SPECK	/ʃpɛk/	
bacon	/ʃ/	[STRIDENT]
	/p/	[CONS] [LAB]
	/ɛ/	[RTR]
	/k/	[CONS] [DORSAL]

The conversion of the speech signal to phonological features is performed in two steps. The task of the acoustic front-end described here is (a) to remove linguistically irrelevant information, (b) to use speaker independent acoustic characteristics to compute the features, and (c) not to exclude potential word candidates due to computational faults or poor signal quality. The general design principle of the system is to use simple and only rough measures that cooperate to form a stable system.

First the signal undergoes a spectral analysis that delivers LPC formants and some rough spectral shape parameters computed from the speech signal using a 20 ms window with 1 ms step rate. The output is an 'online' stream of spectral data as shown in the second panel of Figure 5 (only the speech signal and formant tracks are shown in panels 1 and 2).

The spectral parameters are converted by simple logical decisions into phonological features. The intention is to derive a representation of the speech signal that is relatively independent of the speaker and acoustic line properties. Only very broad acoustic characteristics define the 12 phonological features we use (CONSONANTAL, HIGH, LOW, RTR, VOICE, etc.). For example, the feature [HIGH] is defined by the condition that F1 has to be below 450 Hz. It can be the case that parts of the speech signal incidentally meet or miss the criteria for a particular feature. That is, a non-[LOW] sound that is *not* classified as [HIGH] in its lexical representation might have an F1 below 450 Hz, and another sound segment that should be classified as [HIGH] might have an F1 above 450 Hz. But most important, a sound segment that is [LOW] should *not* have an F1 below 450 Hz. In other words, there is a limit to define a member of a feature undoubtfully, but there is a certain range that does not exclude possible members. In general, however, the acoustic characteristics are chosen so that all members of a particular feature are captured and other sounds might be included as well, but members are hardly missed. And, more important, sounds belonging to a mismatching feature are not captured. The matching conditions and the lexicon eliminate unlikely candidates later. The rationale behind this very relaxed procedure is that in running speech a speaker can deviate from any 'norm' of acoustic characteristics of a sound due to assimilation, coarticulation, dialect, vocal tract parameters, and others. The FUL system does not have such a 'norm'. The system only expects that the feature [HIGH] is acoustically characterised by a low first formant.

The conversion from the spectral characteristics of the speech signal into phonological features delivers a stream of features. Features can change every millisecond as a consequence of the window step rate. Features are defined independently from each other and, hence, they can change independently from other features. For this reason, the features are filtered and time aligned within roughly 20 ms to define feature bundles. These bundles of distinctive features extracted from the speech signal are now compared to those sets stored in the lexicon. This comparison is executed only when the computed feature set changes (and not every millisecond), and the matching logic generates *match*, *no-mismatch* and *mismatch* conditions. The ternary logic works in the same way we have discussed before.

The computation of the matching features relative to the number of features computed from the signal and the number of features stored in the lexicon using the formula given in (3) adds a score for each feature bundle computed from the signal for each entry in the lexicon. The scoring of the consecutive feature bundles gives the word score and its ranking in the list of possible candidates. Feature sets at the beginning of a word gain a higher weight than non-initial features sets; the weight is computed by an exponential decaying function. The set of all word candidates is the lexical cohort that is used to generate word hypotheses.

To recapitulate, all word candidates that match with the initial feature set are activated, together with their phonological, morphological, syntactic, and other information. No segmentation or grouping into syllable units is performed. Matching features increase the scoring for potential word candidates, non-mismatch-

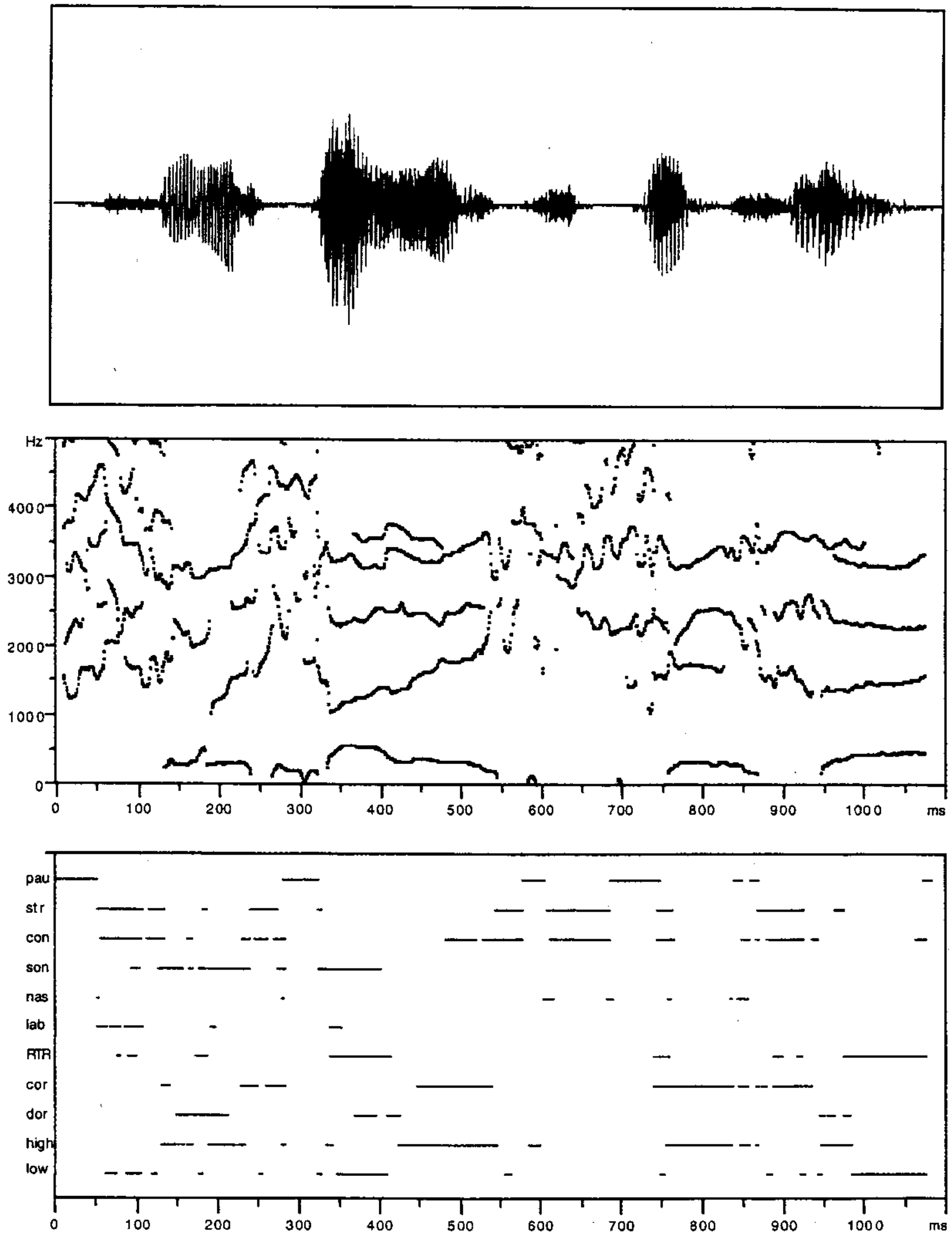


Figure 5: Speech signal, formant tracts and (uncorrected) feature tracks of the sentence "Fußball ist Spitze" (*football is fantastic*), spoken by a male German speaker.

ing features do not exclude candidates, and only mismatching features lead to the rejection of word candidates. The lexicon contains segmental, morphological, semantic, and other information for each word, but for the comparison with the information computed from the acoustic front-end only their representation by phonological features is used. These other information sources are not used to find word candidates in the lexicon but are used to exclude unlikely candidates on a higher level of processing. Characteristic of the system is the operation of these 'higher' level modules in parallel to the acoustic front-end and the lexical access. These 'higher' levels of processing are not described in this paper, which restricts itself to the description of the speech analysis, matching condition and the word hypotheses formation.

For example, the initial feature set [CONSONANTAL][LABIAL][NASAL] activates not only all words beginning with an [m], but also words beginning with other labials that do not mismatch with a nasal (like [b])[10] and also [n] because it is unspecified for place; the ranking of [m] would be higher than [n] [p] [pf] > [b] [f] > [t] [v] > [d]. But if the signal gives [CONSONANTAL][CORONAL][NASAL], a much smaller set is encountered since all dorsal and labial consonants mismatch. The consecutively incoming feature sets deactivate word candidates from the cohort that have mismatching feature sets. In other words, the system overgenerates possible word candidates but does not include impossible word candidates. If the signal gives [HIGH][SONORANT], all high and mid vowels would be activated but *no* low vowels. The rationale behind this mechanism is to include possible variants of sounds (e.g. the vowel /a/ could be pronounced as an [ɔ] or even as [e]) but to exclude variants that will not occur (e.g. the vowel /a/ is never produced as an [i]).

Further, at each point, whenever a word candidate is identified, another new word candidate can begin. Thus, the assumption being that although the signal does not dependably have information of word beginnings or word endings, the lexicon initiates candidates as it goes on.

The main aim of the system is to investigate whether an underspecified representation is suitable to model the linguistic behaviour of humans and their representation of speech. An appropriate evaluation would be therefore a comparison of the system to humans' behaviour. This is beyond our capabilities today with respect to the state of the implementation of the system and to detailed comparable data we have about humans' perception. On the other hand, Hidden Markov Models (HMMs), the standard in automatic speech recognition, operate on different principles and make a direct comparison difficult. HMM systems gain from longer strings of data (states, segments, words, or whatever), because they do not make a definite decision at the smallest unit but delay decisions as long as possible (eventually up to the end of the recognition of a phrase). This is one of the reasons why implausible words might show up in an HMM analysis in the output: the overall probability is maximised even if a part of the string has a very low probability, but there is no 'impossible' label that a part of the string might have. The FUL system makes decisions at the first step, where it rejects candidates about which it is 'sure' that they do not meet a criterion.[11] Thus, this is the first step to compare the FUL system with a HMM system. To make the comparison of single units more compatible, we restrict ourselves to vowels: vowels are more gradient and are more likely to reduce or alter in running speech across speakers and thus allow a more fair comparison.

The Kiel Corpus of Spontaneous Speech (IPDS 1995) served as the database for the comparison. This corpus contains high-quality recordings of spontaneous dialogues of two speakers at a time who were asked to arrange appointments with each other. A total of 54 minutes of speech was recorded for 26 speakers (16 male and 10 female, mostly students from north Germany). The speech data was labelled and transcribed by trained phoneticians. The analysis is based on what the transcribers heard rather than what the speakers intended to say. The Kiel corpus transcription uses 17 vowels in German (all monophthongs, including long-short and tense-lax vowels: [iː, ɪ, yː, ʏ, eː, øː, ɛː, ɛ, œ, aː, a, ɔ, oː, uː, ʊ, ə, ɐ]). For the comparison these were mapped to the 13 vowels the FUL system uses for German since the FUL system does not distinguish long and short vowels on the level of features, but only by moraic representations. Moreover, there is no featural difference in the representation of [a] and [ɐ], and [e] and [ə] which are also only moraically distinguished. Even if we disregard the moraic representation, these simplifications do not lead to a noticeable increase of existing homophones in the lexicon. The complete set of vowels used by both systems is [i, ɪ, y, ʏ, e, ø, ɛ, œ, a, ɔ, o, u, ʊ].

The hidden Markov model had three states and eight mixtures to model every phone;[12] i.e. the system was trained to model the phone and the left and right transitions of that phone (these are the three states) and allowed 8 'variations' of a phone to exist, that are realised by mixtures of Gaussian probability density functions (these are the 8 mixtures of the three states).[13] The phones were modelled left-to-right and no states were skipped. The transformation from the speech waveform to the states was done with 12 MFCC (mel-frequency cepstral coefficients) plus the energy parameters and the corresponding delta-values, giving a total of 26 parameters (cf. e.g. Jelinek, 1997; De Mori, 1998; or Becchetti & Ricotti, 1999 for details about the parameters of stochastic ASR systems). The training of the system was done with a jack-knife procedure, where a subset of the recordings served as training set (i.e. were used to define the pattern sequences for the phones) and another subset (other speakers and other sentences) were used as test set (i.e., had to be 'recognised'). About 80% of the data served as training set and the remaining 20% were used as test set. This procedure was repeated 5 times with different subsets of speakers and sentences selected from the database (i.e. each data set was exactly once in the test set) and the recognition results are averaged over these experiments.

For FUL, we have used only 20 ms of the centre part of the vowels for this comparison. The vowels are classified by combinations of 7 features ([SONORANT], [LABIAL], [CORONAL], [DORSAL], [LOW], [HIGH], [RTR]) and the ternary logic described earlier. Recall, that the FUL system does *not* require any training and therefore there is no separation between training and test sets. Our results are based on a single run.

For both systems, only the top-scoring vowels were counted as 'correct' recognition, i.e., only the vowel(s) with the highest rank were compared to the transcribed phone and counted as correct if they were identical. Note that lower scoring vowels are still contributing to the recognition, both in the HMM and in the FUL system. The results are presented in the Figure 6.

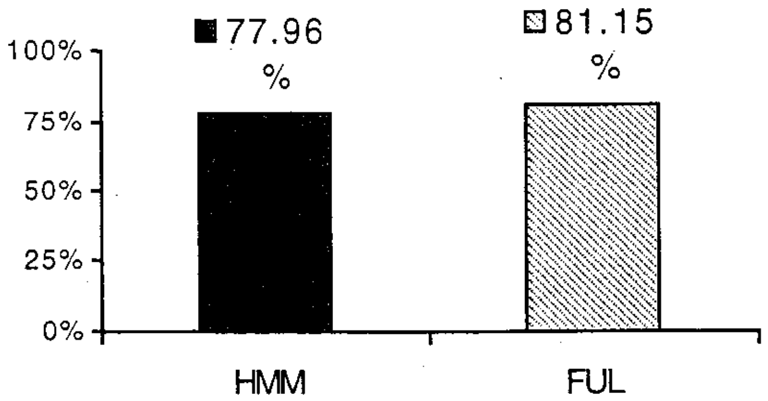


Figure 6: Vowel recognition: HMM and FUL

For the HMM the top-scoring vowels reach 77.96% correct recognition. For the same data set the FUL system achieved 81.15% correct recognition. From these results it seems that the FUL system is able to hold its own in an evaluation format prescribed by stochastic models.

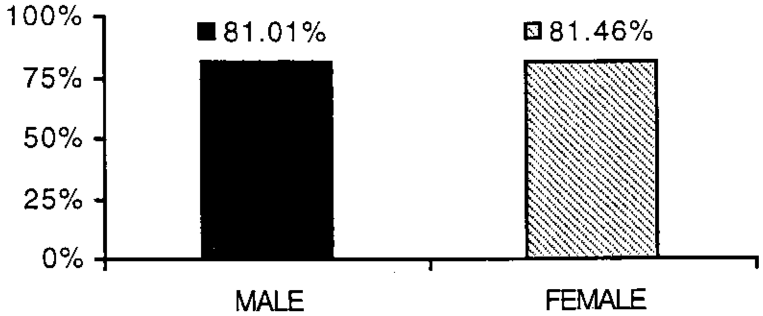


Figure 7: FUL: Vowel recognition by gender

Formants are relatively invariant to spectral tilt, random noise, and overall signal level, which are altered by microphone and transmission-line conditions, hence, these factors do not influence the performance. But formant values do depend on vocal tract size which differs between male and female speakers. For the speakers examined here, the average F1 was 483 Hz for the male speakers and 576 Hz for the female speakers across all vowels, indicating a shorter lip–glottis distance for the female speakers. We therefore examined the differences between the 26 male and 10 female speakers, and found that the vowels were equally well identified across gender. The results are graphically presented in Figure 7. Although there is a very slight bias in favour of the female speakers, the difference is not significant.

In sum, the following characteristics are salient in our implementation of the FUL model in a speech recognition system. It is speaker independent and to a large extent independent of microphone and transmission-line conditions. No training is required, and last but not least, the system is adaptable to other languages because the lexical representation is based on the phonological systems of individual languages.

4.2 Comparable existing models

We now turn to a comparison of our system with existing models which also take recourse to features or related linguistic units, the closest of which are the ACOUSTIC LANDMARK model (Stevens, Perkell & Shattuck-Hufnagel, 1997) and TIME MAP PHONOLOGY (Carson-Berndsen, 1998).

In spirit, the closest model is that proposed by Stevens and his colleagues (1997; earlier LAFF, Stevens, 1992; Stevens et al., 1992), which is also discussed at length in Klatt (1989).[14] The system resembles a more advanced version of the original analysis-by-synthesis principle (Stevens, 1960): A spectrogram is analysed for acoustic characteristics and phonetic segments that relate to these characteristics are proposed as possible candidates. The features associated with these segments are looked up from a table and possible assimilations from neighbouring segments are predicted.[15] The spectrogram is inspected again for acoustic characteristics of these hypothesised features and segment candidates are verified or falsified on basis of this detailed acoustic information. The most prominent acoustic characteristics are hypothesised at segment boundaries (e.g. dropping or missing F1 at VCV boundaries or down-glides of F1 in V-glide-V sequences). The crux of the approach is to examine the speech signal for detailed acoustic characteristics and essentially look for characteristics that might relate to the proposed features. The biggest difference between ACOUSTIC LANDMARK and FUL is that in the former, there is a conversion of the acoustic characteristics into a segment and then features are looked up and searched for. The whole process is not lexicon-driven as the FUL system and is motivated by the acoustical effects at segment boundaries (hence, ACOUSTIC LANDMARKS). The idea behind the system is the handling of allophonic variation after the recognising of the segmental context, whereas the FUL system does neither look for segment boundaries (rather tries to ignore their effects) and handles variations by the matching logic.

Since the system is not implemented by an automatic procedure, it is difficult to compare its performance with FUL or an HMM model.

Another system that converts speech signals into acoustic events and uses them to access the lexicon is most completely described by Carson-Berndsen (1998). The system incorporates two components, the HEAP 'acoustic event' classifier and the SILPA 'phoneme event' recognition module. These two components are described now in more detail.[16]

The HEAP system (Hübener & Carson-Berndsen, 1994) is essentially a statistical categoriser which classifies the speech signal into 24 (later 27, cf. Carson-Berndsen, 1998: 80) 'acoustic events' (like, 'fricative, noisy, nasal, a-like vowel, mid vowel', etc.).[17] This classification is computed from 30 ms frames with a step-rate of 20 ms that are parameterised with 5 cepstrally smoothed PLP coefficients (perceptual linear predictive coefficients, Hermansky, 1990), log energy, and regression coefficients (total 13 coefficients per frame). This recogniser was trained on automatically labelled data on 180 utterances of a single speaker to classify the signal into the acoustic events. To test the performance of HEAP, 20 additional utterances were classified between 77 % and 98 % correctly for a particular acoustic event.

Because the acoustic events are not synchronously changing with the edges of phonemes, a finite-state parser built up a sequence of 'phonological events' that are in turn used for phoneme recognition. That is, the output of the HEAP classifier (i.e. the acoustic events) are converted by a finite state automaton into 'phonological events'. These are 7 independent 'phonological attributes', each one having several values. For instance, the phonological attribute MANNER includes 'plosive, fricative, nasal, lateral, affricate, vowel-like, diphthong'. The phonological attribute

PLACE includes 'labial, apical, palato-alveolar, velar, palatal, uvular, glottal'.[18] In all there are 31 phonological values. Additionally, all possible onset and coda clusters in German syllables were used to restrict the number of possible phoneme sequences derived from the acoustic events. This parser/automaton includes an "underspecified representation of the syllable", but underspecification is understood here as a method to cluster several phonological segments into one 'phonological event'. In this way, 'underspecification' is understood as a state (or memory) saving task rather than as it is understood in phonology (and in this paper otherwise) as a structure that explains certain processes.

Furthermore, the mapping of the acoustic events computed from the signal onto the constraints of the parser are done in a rather different way than in the FUL system. The SILPA parser operates in the following way: if there are more acoustic events in the signal than a particular node of the finite state network needs, the additional events are ignored and the constraint that this node represents is met. If there are less acoustic events than specified in a node, the network could be parameterised so that more important events for a constraint are weighted higher.

The best empirical evaluation scoring rate given is 66.97% for phonemes in a scheduling task scenario with many speakers and 82 utterances (Carson-Berndsen, 1998: 203).

In sum, the system converts the speech signal into 28 'phonetic events' by statistical means that are in turn converted into 31 different 'phonological events' by a finite state machine. The increase in representational units from the signal to the lexical level itself is contradictary to an 'underspecified' representation and rather it is a generation of a detailed phonetic description. Essentially, the use of terminology here is quite different from the description of the FUL system and the two systems are only superficially similar.

5. Conclusion

Our aim has been to present a model of lexical representation which has significant consequences for various aspects of human behaviour, and which can be computationally implemented for the purposes of machine recognition of speech and the testing of models. A lexicon which is phonologically underspecified is the pivot of the FUL model. Phonological variants of morphemes are not listed, the assumption being that the abstract underspecified representation will subsume any phonetic or phonological variation produced by the speaker. The perceptual system extracts phonological features from the signal and directly maps them on to the lexicon. No other linguistic unit is compiled or extracted at this level. There is no intermediate representation like phoneme or syllable. Incoming phonological features activate word candidates constrained by a ternary matching condition, which in turn are fed directly into the phonological and syntactic parser.

Although morphemes are phonologically underspecified, they have sufficient information to distinguish them from each other - unless of course, they are really homophones. This assumption is directly in contrast to a system which assumes that all variants would be listed. The underspecified representation in the FUL system anticipates that there will be variation, but that the variation is itself constrained even at the level of postlexical phonology.

To illustrate, we can take as an example high coronal vowels which can phonologically reduce/lax/unround in running speech. In words like *Füller* [fʏlər] 'pen', *Fühler* [fylər] 'antenna', *Filler* [fɪlər] 'material for smoothing surfaces' and *vieler* [filər] 'much, many-GEN' the first vowel can be indistinguishable. Depending on phrasal structure, rate of speech, focus, and other factors, all the variants can represent one possible pronunciation of each of the words. That is, underlying /y/ could become [ʏ], [i] or [ɪ]; similarly, /ʏ/ could be realised [ɪ], [y] or [i] and so on. All speakers may not have all of the possible pronunciations for all four words, but across speakers it is possible to obtain all variants. Storing the variants makes it impossible to distinguish one from the other; all variants will have equal status unless there is a weighting for each possibility. If this weighting depends on the statistical distribution, the weights depend on the particular data set and it is possible that two of the three variants would have similar weights. FUL predicts a different hierarchy for each variant. If *Filler* was the mispronounced variant of either *Füller*, *vieler*, or *Fühler*, neither word would be a mismatch, but the scores are different as we can see below.

(14) Scores for [ɪ] of *Filler*

Lexical features			Input features of [ɪ]	Score
[y]	*Fühler*	[HIGH, LAB]	[HIGH, COR, RTR]	$1^2/(3 \times 2) = 0.16$
[i]	*vieler*	[HIGH]		$1^2/(3 \times 1) = 0.33$
[ʏ]	*Füller*	[HIGH, LAB, RTR]		$2^2/(3 \times 3) = 0.44$
[ɪ]	*Filler*	[HIGH, RTR]		$2^2/(3 \times 2) = 0.66$

Clearly, when [ɪ] is the surface variant, *Filler* has the highest score. Next in line is *Füller* followed by *vieler* and then finally the last choice would be *Fühler.* The FUL system predicts that for the listener, when [ɪ] is heard, [ʏ] is a better match than [i]. That is, maintaining the laxing (i.e. [RTR]) is preferred. It is entirely possible that storing all the variants with weights would give the same results, but this would have to be done for each lexical item individually. This is not the case for FUL. The predictions would hold for the entire lexicon and would be borne out as a consequence of the underspecified representation and the scoring which incorporates the features extracted from the signal, the features in the lexicon and the matching features.

Since the claim is that FUL models human perception, evidence from both language comprehension experiments and language change were put forward. Language comprehension experiments have shown that listeners extract certain acoustic characteristics reliably, but do not match acoustic details with the lexicon. Rather, the experimental results are best explained with the assumption that lexical access involves mapping of the acoustic signal to an underspecified featural representation such that non-mismatching variants are treated differently from mismatching variants. If an underspecified representation is indeed part of the adult mental lexicon, then one assumes that the language learner is able to construct such a representation. If so, then language change ought to provide evidence that an underspecified representation at a certain point of time lends itself to a different pattern of change than a more fully specified representation. Our example came from the change of geminates to nongeminates from Latin to the northern Italian dialect of Miogliola. Although degemination occurred everywhere, the original length contrast in Latin obstruents could be maintained by spirantising the original nongeminate stops. This was not possible for sonorants where in general the contrast was lost, the exception being the coronal nasals. For these consonants, the original underspecified representations were exploited, such that the Latin nongeminate /n/ remained underspecified in Miogliola, but the geminate /n:/ degeminated but acquired a place feature.

The computational adequacy of these assumptions was verified in implementing an automatic speech recognition system. Again, assuming that speech is variable but that the variation is constrained, the FUL ASR system focuses on solving the problem of recognition not by capturing all possible details from the signal but by extracting acoustic characteristics which can be easily interpreted as distinctive features which are relevant for distinguishing lexical representations. The information retrieved from the signal is not responsible for building lexical representations. The lexical representations in their idealised forms already exist and the information from the signal (i.e. the extracted features) is mapped onto existing representations. Resolving variation is achieved by the fact that given underspecification and the ternary matching logic, a one-to-many matching is possible. Since particular underspecified representations are geared towards accepting only phonologically viable phonetic variants, the one-to-many matching is not random.

We should add at this point, that our main objective is not to construct the most marketable speech recognition system. Product-oriented systems have specific constraints and individual requirements. In principle, the FUL system is adaptable for specific products, but this has not been our main concern. A system to reconfirm flights, for instance, does not require a complex model like FUL. A limited vocabulary combined with an intelligent dialogue is a far better solution. Our aim has been to construct a computational system which operates on the principles we believe are important for human perception. We would like to make it entirely speaker independent and not use any stochastic procedures, thereby no doubt sacrificing possible gains. However, as it stands, FUL can provide a means of testing speech perception theories, particularly details of feature interaction, properties of features, lexical representations, coarticulation and such. It could also be an excellent tool to study dialect variations and possible directions of change. Since FUL takes the speech data as speech and not as any random acoustic signal, and assumes that this speech is produced by speakers who have a real language in their heads, it is intended primarily as a linguistic tool, using linguistic primitives and exploiting linguistic knowledge.

The FUL system is highly constrained – in the allowable lexical representations, in what is extracted from the signal, and in the information used to make the matching decisions. The message has been "Less is More" in a positive sense.

Acknowledgements

We are especially grateful to John Kingston and an anonymous reviewer's very detailed comments. We are much indebted to Jacques Koreman for the HMM classifications. This paper also owes a large debt to Jennifer Fitzpatrick, Sibrand van Coillie and Michael Wagner. The research was partially supported by the SFB 471, the Max-Planck-Forschungspreis and the Leibniz Prize.

Notes

[*] We would like to dedicate this paper to Mirco Ghini †, without whose abounding enthusiasm and intellectual commitment, a large part of this research would have never been possible.

[1] There are other proposals outside phonological approaches, like 'full listing models' that abandon generalisations altogether, or 'exemplar models' (e.g. Medin & Schaffer, 1978; Nosofsky, 1986) that use individual items as representatives for a category, or 'prototype models' (e.g. Klatt, 1979) that compute an average representative for a category. There appear to be misconceptions regarding the terminology. For instance, Bybee (2000: 253) refers to her model of the lexicon as an 'exemplar model', while assuming that "Each experience of a word is stored in memory with other examples of use of the same word". Such an assumption fits with a 'full listing model'. To cover all these models and their variants would go far beyond the scope of this article.

[2] For support for underspecification see, for instance, Keating (1988), Kiparsky (1993), and Rice (1996); psycholinguistic evidence is provided in Lahiri, Jongman & Sereno (1990), Lahiri (1991), Lahiri & Marslen-Wilson (1991, 1992), and Fitzpatrick & Wheeldon (2001). Opposing views have been presented, for example, in McCarthy & Taub (1992), Mohanan (1993) and Steriade (1995), and references therein.

[3] Mohanan (1993) gives a hierarchy of assimilation possibilities where the most frequent type is coronal assimilation. In languages where labials do assimilate to other places of articulation, a dental/alveolar sound is always subject to assimilation. This persuades Mohanan to assume that there is no underspecification but rather a hierarchy of 'attraction'. In our model, for the labials the assimilation must be a result of delinking-cum-spreading and would be treated differently from coronal assimilation.

[4] Full statistical and methodological details are not repeated here since this is an overview and the original paper is being published in an experimental journal. A total of 24 monosyllabic words with final /-n/ and 12 monosyllabic words with final /-m/ were used as primes, each with a semantically related target (e.g. *Bahn-Zug*). The differences in the number of items was due to the fact that there were less words in the language ending with /-m/ where the final consonant could be changed to make nonwords. A total of 144 German native speakers were tested.

[5] Significance was tested at a 5% level.

[6] There were 20 words each with medial /n/ and /m/. A total of 90 subjects participated in this experiment.

[7] For a recent review of the different predictions and experimental evidence for lexical access based on underspecification or full specification, see Fitzpatrick & Wheeldon (2001).

[8] Coronal stops also became affricates, but this is not important for the discussion here.

[9] The total number of features needed may be language dependent.

[10] The signal feature [ABRUPT] mismatches with the lexical feature [NASAL], cf. (7a). If [NASAL] is found in the signal it cannot mismatch with [ABRUPT] because [ABRUPT] is not stored in the lexicon.

[11] Note, however, that this is a much more relaxed decision than in many early phonetics-based systems, that tried to determine the set of possible segments.

[12] The HMM experiments were run at the University of Saarbrücken by William Barry, Jacques Koreman and their colleagues.

[13] To use the left and right context in modelling phones to allow different contexts in ASR was already proposed by Klatt (1979).

[14] At a workshop at Schloss Freudental (Konstanz) July 1998 entitled *Speech Recognition: Man and Machine*, Ken Stevens and his colleagues presented the system in detail.

[15] The original texts do not make a clear distinction between acoustic characteristics and the phonological features as it is presented here. Both are understood by Stevens and his colleagues as different expressions of the same thing.

[16] The 'acoustic events' are very different from phonological features and neither are they acoustic characteristics, as in Stevens' model.

[17] The classifier for the acoustic events was originally planned as a deterministic module that uses auditory spectra as input (Hübener, 1991).

[18] Overall, the system maps the acoustic events to phonetic descriptions rather than phonological features.

References

Becchetti, C. & Ricotti, L. P.
1999 *Speech Recognition - Theory and C++ Implementation.* Chichester: John Wiley & Sons.

Bybee, J.
2000 Lexicalization of sound change and alternating environments. M. B. Broe & J. B. Pierrehumbert (eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, (pp. 250–268). Cambridge: Cambridge University Press.

Carson-Berndsen, J.
1998 *Time Map Phonology.* Dordrecht: Kluwer.

De Mori, R.
1998 *Spoken Dialogues with Computers.* London: Academic Press.

Fitzpatrick, J. & Wheeldon, L.
2001 Phonology and phonetics in psycholinguistic models of speech perception. In N. Burton-Roberts, P. Carr, & G. J. Docherty (eds.), *Phonological Knowledge – Conceptual and Empirical Issues*, (pp. 131–160). New York: Oxford University Press.

Gaskell, G. & Marslen-Wilson, W. D.
1996 Phonological variation and inference in lexical access, *Journal of Experimental Psychology: Human Perception and Performance*, **22**, 144–158.

Ghini, M.
2001 *Asymmetries in the Phonology of Miogliola.* Berlin: Mouton.

Ghini, M.
to appear The role of underspecification in the development of metrical systems. In P. Fikkert & H. Jacobs (eds.), *Change in Prosodic Systems.* Berlin: Mouton.

Gow, D.
2001 Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, **44**, 1–27.

Hermansky, H.
1990 Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America,* **87**, 1738–1752.

Hübener, K.
1991 *Eine Architektur zur integrierten Analyse von Sprachsignalen.* Verbundprojekt ASL, Document Nr. ASL-TR-3-91/UHH, Vers. 2. University of Hamburg.

Hübener, K. & Carson-Berndsen, J.
1994 Phoneme recognition using acoustic events, *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP 94),* Yokohama, **4**, 1919–1922.

IPDS
1995 *The Kiel Corpus of Spontaneous Speech, Vol. 1.* CD-ROM. Kiel: Institut für Phonetik und digitale Sprachverarbeitung.

Jelinek, F.
1997 *Statistical Methods for Speech Recognition.* Cambridge: MIT Press.

Keating, P. A.
1988 Underspecification in phonetics. *Phonology,* **5**, 275–292.

Kiparsky, P.
1993 Blocking in nonderived environments. In S. Hargus & E. M. Kaisse (eds.), *Phonetics and Phonology,* (pp. 277–314). New York: Academic Press.

Klatt, D. H.
1979 Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics,* **7**, 279–312.

Klatt, D. H.
1989 Review of selected models of speech perception. In W.Marslen-Wilson (ed.), *Lexical Representation and Process,* 169–226. Cambridge: MIT Press.

Lahiri, A.
1991 Anteriority in sibilants. *Proceedings of the XIIth International Congress of Phonetic Sciences,* Aix-en-Provence, **1**, 384–388.

Lahiri, A.
1999 Speech recognition with phonological features. *Proceedings of The XIVth International Congress of Phonetic Sciences,* San Francisco, **1**, 715–718.

Lahiri, A.
2000 Phonology: structure, representation and process. In Linda Wheeldon (ed.), *Aspects of Language Production,* (pp. 165–225). Hove: Psychology Press.

Lahiri, A. & Coillie, S. van
to appear Non-mismatching features in language comprehension.

Lahiri, A. & Evers, V.
1991 Palatalization and coronality. In Paradis, C. & Prunet, J.-F. (eds.), *The Special Status of Coronals,* (pp. 79–100). San Diego: Academic Press.

Lahiri, A., Jongman, A. & Sereno, J. A.
1990 The pronominal clitic [dər] in Dutch: a theoretical and experimental approach. *Yearbook of Morphology,* **3**, 115–127.

Lahiri, A. & Marslen-Wilson, W. D.
1991 The mental representation of lexical form: a phonological approach to the recognition lexicon. *Cognition.* **38**, 245–294.

Lahiri, A. & Marslen-Wilson, W. D.
1992 Lexical processing and phonological representation. In G. J. Docherty & D. R. Ladd (eds.) *Papers in Laboratory Phonology II: Gesture, Segment, Prosody,* (pp. 229–254). Cambridge: Cambridge University Press.

McCarthy, J. & Taub, A.
1992 Review of Paradis and Prunet 1991. *Phonology,* **9**, 363–370.

Medin, D. L. & Schaffer, M. M.
1978 Context theory of classification learning. *Psychological Review,* **85**, 207–238.

Mohanan, K. P.
1993 Fields of attraction in phonology. In John Goldsmith (ed.) *The last Phonological Rule: Reflections on Constraints and Derivations*, (pp. 61–116). Chicago: University of Chicago Press.

Nosofsky, R. M.
1986 Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General,* **115**, 39–57.

Reetz, H.
1998 *Automatic Speech Recognition with Features.* Habilitationsschrift. Universität des Saarlandes.

Reetz, H.
1999 Converting speech signals to phonological features. *Proceedings of The XIVth International Congress of Phonetic Sciences*, San Francisco, 3, 1733–1736.

Rice, K.
1996 Default variability: the coronal-velar relationship. *Natural Language and Linguistic Theory,* **14**, 493–543.

Steriade, D.
1995 Underspecification and markedness. In J. Goldsmith (ed.) *The Handbook of Phonological Theory,* (pp. 114–174). Cambridge: Blackwell.

Stevens, K. N.
1960 Towards a model for speech perception. *Journal of the Acoustical Society of America,* **32**, 47–55.

Stevens, K. N.
1992 Lexical access from features. *Speech Communication Group Working Papers, Research Laboratory of Electronics, MIT,* **8**, 119–144.

Stevens, K. N., Manuel, S. V., Shattuck-Hufnagel, S. & Liu, S.
1992 Implementation of a model for lexical access based on features. *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP 92),* Banff, **1**, 499–502.

Stevens, K. N., Perkell, J. S. & Shattuck-Hufnagel, S.
1997 Speech Communication. *MIT-Research Laboratory for Electronics Progress Report,* **140**, 353–367.

Laboratory Phonology 7

edited by

Carlos Gussenhoven
Natasha Warner

Mouton de Gruyter
Berlin · New York 2002