UNDERSPECIFIED PHONOLOGICAL FEATURES FOR LEXICAL ACCESS

Henning Reetz

Department of Linguistics, University of Konstanz, Germany
henning.reetz@uni-konstanz.de, http://www.ling.uni-konstanz.de/pages/home/reetz

Abstract

The FUL (featurally underspecified lexicon) system of automatic speech recognition is based on the representation of words in the lexicon with underspecified distinctive features. The speech signal is converted from the waveform into an online spectral representation made up of LPC formants and a few parameters describing the overall spectral shape. These spectral parameters are converted into distinctive phonological features by simple logical decisions. The phonological features, in turn, are compared with all entries in the lexicon. No classification into phone(me) segments, syllables, or spectral templates is used for the selection of words from the lexicon. Comparison of signal features with those stored in the lexicon uses a ternary system of matching, no-mismatching, and mismatching features. Matching features increase the scoring for potential word candidates, no-mismatching features do not exclude candidates and only mismatching features lead to the rejection of word candidates. Activated (i.e., not rejected) word candidates are expanded to include word hypotheses, even without further acoustic evidence, and are used in the phonological and syntactic parsing that operates in parallel with the acoustic front-end. Only after the activation of a word from the lexicon as a possible candidate additional information about it, like prosodic, morphological, syntactic, and, ideally, semantic information becomes available.

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, 161-173.

1. Theory

The speech signal for the same phonetic segment varies across dialects and speakers, within speakers in certain segmental and prosodic contexts, and even for the same speaker and context with repetition, speaking rate, emotional state, microphone, etc. Not surprisingly, speech recognition with simple spectral template matching has limitations. Any variation in the signal leads to variation of the spectra that are compared to the stored templates. Statistical approaches like Hidden Markov Models based on large training sets have led to acceptable results, but are still speaker and transmission line dependent or operate only with a restricted vocabulary, syntax, and semantics.

The FUL system (Reetz, 1998) operates on a completely different principle. No spectral templates are computed from the speech signal to access the lexicon, nor is the signal analysed in great detail for acoustic evidence of individual segments and their boundaries. Instead, the signal is converted into speaker independent sets of phonological features. These features are compared with the feature sets stored in the lexicon using a ternary logic (see Section 2.1). The task of the acoustic front-end described in Section 2.3 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.

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. The extracted features are compared with the stored feature sets of the 50000 base form word lexicon whenever the set of computed features changes. This 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. That is, a word like “bean” is represented with three slots of feature sets:

/	b	i	n	/
	[cons]	[vocalic]	[cons]	
	[labial]		[nasal]	
	[voice]			
		[high]		

The comparison between the features from the signal and the features stored in the lexicon uses a ternary logic that is described in Section 2.1. This process handles within and across word assimilations and can deal with (dialectal) allophonic variants as long as a feature does not change to an opposite category (e.g. a [front] vowel becomes a [back] vowel). The entire word is activated as a possible candidate if the feature sets of the initial part of a word do not mismatch with the feature sets computed from the signal. Furthermore, morphosyntactic variants of the activated words are generated by rule without acoustic evidence. This process can handle certain reductions and deletions that are common in fluent speech. Additionally, the initial part of a word obtains a higher weighting than the later parts and compounds are mostly stored as individual words with appropriate markers.

To repeat, acoustic evidence is transformed into a set of speaker independent distinctive features that activate word candidates. These word candidates generate word hypotheses without specific acoustic evidence. The subsequently extracted sets of distinctive features reject mismatching word candidates. In addition, if the sequence of already processed feature sets complies with at least one complete word candidate, another set of new word candidates is activated.

All word hypotheses together with their stored linguistic information are fed into the syntactic and phonological parser that uses additional prosodic and other information. The 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. These ‘higher’ level modules of the system operate in parallel with the acoustic front-end and lexical access. That is, the system does not wait until a possible word is recognized by the acoustic front-end, but operates with incomplete acoustic information. These ‘higher’ levels of processing are not described in this paper, which restricts itself to the description of the acoustic front-end, the matching process, and the word hypotheses formation.

The general design principle of the system is to use simple and only rough measures that in concert form a stable system. This behaviour could be compared to beavers building a dam – they use many roughly cut stems of trees und combine them in such a way that they form a robust dam. The goal is to be able to handle massively faulty data rather than to expect clean data with a high signal-to-noise ratio. The idea is to mimic the behaviour of human listeners, who seem to be unconcerned by adverse acoustic conditions and are able to resolve a wide range of variations and assimilations, with apparent ease.

1.1. *Motivation: assimilations in running speech*

Ambiguities in the signal, whether they come from random noise or whether they are linguistic in nature, like cliticisations of words or assimilations, are the norm rather than the exception in natural language. Human listeners, however, appear not to be worried by adverse acoustic conditions and handle variations in the signal with ease. Language comprehension experiments (Lahiri & Marslen-Wilson, 1992; Lahiri et al., 1990) have shown that listeners extract certain acoustic characteristics, but do not match acoustic details with the lexicon. Rather, the experimental results are best explained with the assumption that lexical access involves mapping the acoustic signal to an underspecified featural representation.

For example, the 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 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”). Simple articulatory mechanics cannot account for this, because an articulatory assimilation would operate in both directions. An explanation can be given by assuming that coronal sounds are underspecified for place, whereas labial and dorsals are not: the labial place of articulation spreads to the preceding coronal sound (if the language has regressive assimilation) which 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 labials or dorsals cannot change their place.

This explanation is straightforward for speech production, but what about speech perception? How can a realisation of “gree[m]” in a labial context (like “green bag”) or “gree[ŋ]” in a dorsal context (like “green grass”) lead to the access of the word “green” in the lexicon? Normally, “gree[m]” and “gree[ŋ]” are nonwords in English. And, at the same time, how should a mechanism be constructed to allow the activation of the word “bean” as well as “beam” if the acoustic input is “bea[m]” (as in “bean be”), when “bean” is a word of the language? Human listeners handle these asymmetries (and many other assimilatory effects) within and across words without noticing them, as reaction time experiments have shown (Lahiri, 1995). The solution to these seemingly contradictory requirements can be obtained (i) 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 (ii) by postulating a ternary matching logic in the signal-to-lexical mapping.

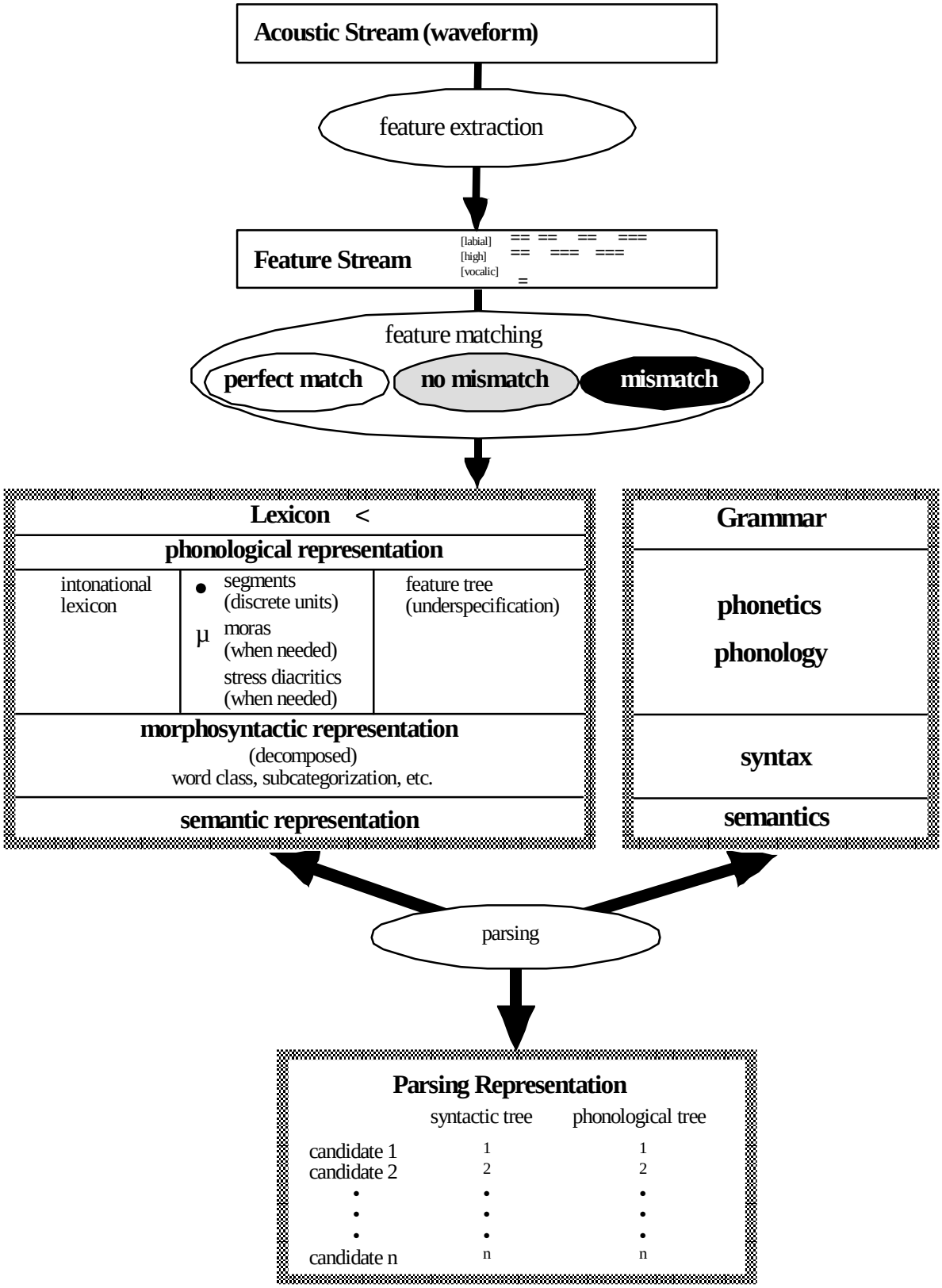

Figure 1. General layout of the FUL system

2. System description

Figure 1 displays the general layout of the FUL system. The speech signal is converted into distinctive phonological features. The subsequent text first describes the matching process because it is the central part of the system, then the morphological extension is motivated and described. Finally, the acoustic front-end is described in more detail.

2.1. *The ternary matching logic*

The distinctive features extracted from the speech signal are compared to those sets stored in the lexicon. The feature sets are computed every millisecond from the signal and are compared to the lexicon whenever the computed feature set changes. (The construction of the signal-to-feature conversion ensures that individual features change synchronously, i.e., transitional states from feature set ‘A’ to feature set ‘B’ are removed by this logic; see Section 2.3.3 for details.) The matching logic generates *match*, *no mismatch*, and *mismatch* conditions that are explained now.

The *match* condition can only occur if both the signal and the lexicon have the same features. This condition is used for scoring the word candidates and includes a correction formula to account for feature sets of different sizes.

Mismatch occurs if the signal and the lexicon have contradictory features. A mismatch excludes a word from the list of possible word candidates. Mismatching features can be either contradictory in both directions (e.g. [high] and [low] mismatch, independent of which is extracted from the signal and which is stored in the lexicon) or they can be underspecified in the lexicon but are extracted from the signal. 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 stored features [labial] and [dorsal]; for example, if the features extracted from the signal conform to the sequence of segments [bin], the word candidate /bim/ (“beam”) is rejected, as explained in Section 1.1.

A *no mismatch* situation 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. Case (i), when no feature is computed 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, but neither does the matching score of a candidate

increase. Case (ii) is exactly the case for lexical features like [coronal] or [abrupt], neither of which are 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 as stored in the lexicon. That is, the signal [grim] *does not* deactivate /grin/ (“green”). Coronals get simply a lower score than the labials (or dorsals), which obtain a match, but coronals are not excluded. They remain active as assimilatory variants.

The system counts the number of matching features for each frame, i.e., each time a feature set computed from the signal has changed. The scoring of the consecutive frames gives the word score and its ranking in the list of possible candidates. Feature sets at the beginning of a word gain a higher weighting where the weight is computed by an exponentially decaying function. The set of all word candidates is the lexical cohort that is used to generate word hypotheses.

2.2. *The lexical cohort*

As soon as they are available, the features that are extracted from the signal are compared to all entries in the lexicon. No segmentation or grouping into syllable units is performed. All word candidates that match with the initial feature set are activated, together with their phonological, morphological, syntactic and other information. That is, an initial feature set [consonantal] [labial] [nasal] not only activates all words beginning with an [m], but also words beginning with [n], [p] and others. 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. The rationale behind this mechanism is to include possible variants of sounds (e.g. the vowel /a/ could be pronounced as an [ɛ] or even as [e]), but to exclude variants that will not occur (e.g. the vowel /a/ is never produced as an [i]).

The activated word candidates that fulfill the acoustic criteria of the signal also generate possible morphosyntactic variants without having acoustic evidence for them. For example, the feature sequence that is appropriate to activate the word “fang” (*catch* IMP.) in German also activates the word forms “fange” (*catch* 1.PRES.SG.), “fangen” (*catch* 1.PRES.PL.), and even the separable verbs such as “an-fangen” (*to begin*) or “um-fangen” (*to encircle*). These words are used to generate syntactic phrase hypotheses without waiting for further acoustic evidence. For example, the acoustic information that conforms to the (partial) phrase “ich fang” (*I catch* IMP.) (which is not grammatical) also generates the sentence hypotheses “ich

fange" (*I catch* 1.PRES.SG.) and "ich fange an" (*I start* 1.PRES.SG.), among others. This has many advantages. For instance, many word final morphological markers in running speech are not pronounced. The three German phrases "Fang an!" IMP.SG." (*start!*), "Ich fang[ə] an." (*I begin.* 1.PRES.SG.) and "Wir fang[ən] an." (*We begin.* 1.PRES.PL.) are usually all pronounced [...faŋ] without any acoustic evidence for the [ə] or [ən]. The recognition system does not need to look for any acoustic trace in the signal for a schwa of "fange/fangen"; rather, it generates these word hypotheses from the available acoustic data. Consequently, these forms of reduction and deletion are handled by the system without storing variants in the lexicon.

2.3. *The acoustic front-end*

The conversion of the speech signal to phonological features is performed in two steps with one correction applied in each step.

First, the signal undergoes a spectral analysis that delivers LPC formants and some rough spectral shape parameters (Section 2.3.1). In the second step the formant and spectral shape data are converted by simple heuristic filters into phonological features (Section 2.3.2). The feature data in turn undergoes a heuristic filter to remove irregularities in the feature streams (Section 2.3.3). Additionally, the individual tracks of features are time aligned so that changes in the feature sets occur in all features at the same time. These steps are described below.

2.3.1. *Spectral analysis*

A standard LPC analysis and a standard FFT power spectrum are computed from 20ms Hamming windowed stretches of speech. The first five formants are taken from the LPC analysis, the overall spectral slope, the spectral slope below, and the slope above 2.5 kHz are computed from the FFT spectra. The window step size is 1 ms. That is, in the time domain a very dense sequence of quite roughly computed spectra is available. The intention is to convert the speech signal into a stream of spectral data. This temporal density might not be necessary for the system at every part of the signal, but it is currently used as redundant information to separate accidental glitches in the data from systematic variations due to properties of the signal.

The outcome of the first step, the spectral analysis, is a set of 5 formants and 3 spectral slope parameters computed every millisecond from a 20 ms window.

2.3.2. *Converting spectral parameters to phonological features*

The spectral parameters are used to compute phonological features. These features should reflect important properties of the speech signal and they should be relatively independent of the speaker and acoustic line properties. It is *not* intended that the features exactly match the features that are stored in the lexicon; only features that are impossible for a certain sound (e.g. a low vowel should never be [high]) should not be computed from the signal. That is, a low vowel should preferably have the feature [low] in the acoustic signal, but a missing [low] computed from the signal does not exclude it from the subsequent processing. Only if the feature [high] is computed from the signal a low vowel is excluded in the lexical access.

The system uses very broad acoustic characteristics to define the phonological features. We investigated two databases (TIMIT, 1993; IPDS, 1995) and found that it is possible to characterise individual segments for individual speakers with more complex acoustic characterisations, but these characteristics could not be extended to different speakers. We also tested some speaker normalisation procedures, but we found a rather broad specification of acoustic characteristics for the phonological features for the raw spectral data more efficient than the application of a speaker normalisation procedure prior to the classification. We observed an improvement by the speaker normalisation procedures with carefully measured formants of speech signals recorded under optimal conditions, but we observed a deterioration with the normalisation procedure with automatically extracted formant data. Furthermore, we used a formant-tracking algorithm to correct for accidental mis-computation of formants. This correction improved the formant trajectories, but led on the other hand to decision problem about the numbering of formants and sometimes stabilised erronous formant computations. In total, we observed a better overall behaviour of the system *without* the corrections.

Most of the 13 phonological features we use (consonantal, vocalic, continuant, RTR, voice, strident, abrupt, nasal, labial, coronal, dorsal, high, low) are defined by simple acoustic characteristics. For example, the feature [high] is defined by $F1 < 450$ Hz. These very simple characteristics are often true for sounds that do not belong to a particular feature (e.g., some other sound that is not [high] might have an F1 below 450 Hz), but crucially this acoustic characteristic seldom misses any sound that is [high]. That is, the acoustic characteristics are chosen so that all members of a particular feature are captured and other sounds might be included as well, but no member is missed. The construction of the matching process and the lexicon (see Section 2.1) eliminates implausible 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’ (for example, as it is set up during the training phase of a Hidden Markov Model). The system only expects that the feature [high] is acoustically characterized by a low first formant and that a high first formant would be identified by a listener as a low vowel.

Certain characteristics can be more complex. For instance, the feature [abrupt] includes an investigation of the change of all formants in a short period of time. The acoustic characterisation of this feature includes a description of the temporal development of all formants, and not simply the sudden increase of energy, as might be expected from a plosion release. The reason is that stops in running speech are not necessarily marked by a nice closure and a sudden release, as it is common in laboratory speech. Stops can become quite reduced to a short disturbance in a speech signal, without a preceding stretch of pause.

We tried to avoid dependencies between features, but some features are dependent on other features. This is not only a consequence of the feature hierarchy, but also a consequence of acoustic dependencies. The number of dependencies due to the feature hierarchy is small, because consonants and vowels share the same features (e.g., the feature [high] is used for consonants as well). But this is a burden for acoustic dependencies. For example, the acoustic specifications for the feature [labial] is differently defined for vowels (with an open oral tract), nasals (with a closed oral tract – German has no nasal vowels), and stops (with a closed vocal tract). These different articulatory conditions lead to different acoustic characteristics that are used in the heuristics to specify the feature [labial]. In particular, [labial] depends on [abrupt] and [nasal].

The outcome of the extraction of features from the spectral data of the speech signal is a stream of phonological features. Like the spectral data, the features are computed every millisecond. Unlike the spectral parameters some features can include a relation to adjacent spectral data, as discussed for the feature [abrupt], which by definition includes the inspection of a change over time.

2.3.3. *Correcting phonological features*

The phonological features computed in Section 2.3.3 are described by acoustic characteristics. Some features inspect the time course of the signal (or its spectral data), but most features are set (or not set) on the instantaneous presence of their

acoustic characterisation. Consequently, features can be present or not at any time and they can change their state independent of other features. This can lead to the existence or non-existence of features for a time period of only one millisecond, which can be due to a noisy signal channel or a fault of the formant or spectral shape computation. It can also be a consequence of the speaker's behaviour. For example, the change from one sound to another normally changes more than one feature. The speaker does not change all these features at the same millisecond. Rather, by the nature of articulation, within a certain period of time, features change one after the other. Actually, some models of automatic speech recognition focus on these periods of changes in the signal (e.g. Stevens et al., 1997), but the FUL system tries to locate more stable areas. This approach is based on the consideration that any noise will lead to many sudden changes in the signal and that the more robust parts in a speech signal must be more salient to be distinguishable from any noise. Furthermore, the entries in a lexicon are defined by a sequence of feature sets. The transition state would be interpreted as a sequence of rapidly changing feature sets. If they would be forwarded to the lexicon, the lexicon would have to store all combinations of transitional states, exactly the opposite of the FUL approach, which wants to keep variation out of the lexicon and handle variation in the matching process during lexical access.

This overview described the task of the correction system for the phonological features: first, spurious changes have to be removed and, second, changing features have to be synchronized. A special treatment is required for the feature [abrupt] because this feature (a) is defined by a very rapid change, which is considered to be an error condition for all other features, and (b) can have a very brief period of acoustic characteristics to define the place of articulation adjacent to it, which would be considered to be a spurious occurrence for all other features. Consequently, the existence of the feature [abrupt] forced a special treatment of all features in the adjacent ±20 ms that is not described in detail here, and it blocks all operations in the other feature tracks that are described below.

For all other features, the algorithm first searches each feature track for gaps that are shorter than 5 ms and fills such gaps by inserting the feature. Second, isolated stretches of features in a track that are shorter than 15 ms are removed. These two corrections are the only corrections currently applied to the feature tracks. Some 'improved' filters that we tested seem to operate well for individual speakers, but applied to all speakers in the same way they seem to decrease the overall performance of the system.

The synchronisation of the individual features is first performed for the feature [abrupt]. All features directly adjacent to the feature [abrupt] that persist for at least

10 ms are set together with the feature [abrupt] for 10 ms, otherwise they are removed. Then, the procedure described below is applied without investigating the other features outside these 10 ms ‘abrupt’ windows.

If any of the features changes its value (i.e. becomes existent or disappears), the next 20 ms are investigated. If no other feature change occurs, the 20 ms are treated as a set of stable features, which are used to access the lexicon. If any feature changes its value during these 20 ms, the period is considered unstable and a new 20 ms search for stability is triggered. Roughly speaking, this procedure leads to sets of features that are at least 20 ms long, but that give rise to only one lexical access. The length of the stability of a feature set could be used as information as well, but in the present implementation this information is disregarded.

3. Summary

In sum, the FUL system has the following crucial characteristics. The lexicon consists of words or rather morphemes whose phonological representation is underspecified. Each word has a unique phonological representation – i.e. no word variants are listed. The speech signal is converted from the waveform into an online spectral representation made up of formants and a few parameters describing the overall spectral shape. These LPC and spectral parameters are converted into distinctive phonological features which, in turn, are directly compared with all entries in the lexicon. No classification into segments, syllables or spectral templates is used for the selection of words from the lexicon. A ternary matching procedure constrains the list of word candidates, which are fed directly into the phonological and syntactic parser. *Matching* features increase the scoring for potential word candidates, *no-mismatching* features do not exclude candidates, and only *mismatching* features lead to the rejection of word candidates. The word candidates are expanded to include word hypotheses, even without complete acoustic evidence. The system is speaker independent and to a large extent independent of microphone and transmission line conditions. No training is required and, last but not the least, the system is adaptable to other languages. What needs to be known for a new language are the phonemic oppositions of the language and the feature set to encode these oppositions. Additionally, the morphology and morphophonological interactions have to be known to construct the list of entries in the lexicon. This sort of information is available from linguistic sources and there is no need to construct large labelled databases to train the words of a language.

The system is implemented as it is described here. Ongoing work investigates alternative methods in the acoustic front-end and extends the morphological expansion. The present results indicate that formants are very robust parameters for many sounds, but that other parameters than the overall spectral shape might be needed to describe all phonological features reliably. The extraction of spectral parameters is tested with methods that improve the estimation of the parameters in the presence of background noise, including simultaneous speakers. Speaker normalisation procedures seem not to improve the system. The conversion from the spectral parameters into phonological features is also tested with algorithms that take more than one feature into account. Furthermore, the system intially used the CELEX database. This database does not represent the words in a form that is ideally suited for this application and the construction of a new lexicon with about 50 000 base forms and rules for morphological expansions (inflections, derivations and compounding including ab- and umlauting) is underway. And, naturally, a comparison with standard HMM-based systems will be performed when the complete lexicon is available.

4. References

CELEX (1993). *The celex Lexical Database.* CD-ROM. Nijmegen: Centre for Lexical Information & Max Planck Institut für Psycholinguistik.

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

Lahiri, A. (1995). Undoing place assimilations (Invited talk at the 129th meeting of the ASA). *J. Acoust. Soc. Am.* **97**, 3333.

Lahiri, A., Jongman, A. & Sereno, J. (1990). The pronominal clitic [´r] in Dutch. *Yearbook of Morphology* **3**, 115-127.

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

Reetz, H. (1998). *Automatic Speech Recognition with Features* (Habilitation thesis, University of the Saarland).

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

TIMIT (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. CD-ROM. LDC.