Syllable-level desynchronisation of phonetic features for speech recognition

Katrin Kirchhoff

Technische Fakultät
University of Bielefeld
33501 Bielefeld
Germany

Abstract

This paper describes a novel approach to speech recognition which is based on phonetic features as basic recognition units and the delayed synchronisation of these features within a higher-level prosodic domain, viz. the syllable. The object of this approach is to avoid a rigid segmentation of the speech signal as it is usually carried out by standard segment-based recognition systems. The architectural setup of the system will be described, as well as evaluation tests carried out on a medium-sized corpus of spontaneous speech (German). Syllable and phoneme recognition results will be given and compared to recognition rates obtained by a standard triphone-based HMM recogniser trained and tested on the same data set.

1. Introduction

A well-known inadequacy of standard stochastic speech recognisers is that they map the speech signal, which consists of parallel, temporally overlapping acoustic properties, to discrete sequential units, thus ignoring coarticulatory effects. Context-dependent recognition units (e.g. triphones) or sequential subsegmental units (microsegments), which are employed in order to achieve better modelling of coarticulation, greatly increase the recognition inventory and thus enhance storage and processing requirements. Several speech recognisers [1, 2, 3, 4, 7, 10, 9] have been designed to overcome these deficiencies by using parallel subsegmental units, notably phonetic features, as primary recognition units. However, these systems often do not allow for the temporal misalignment of units: features are stacked to form discrete vectors which then serve as input to a higher-level classifier [1, 2, 9]. Where featural overlap is taken into account, it is usually confined to phoneme-sized temporal domains ([3, 4, 7]). Since articulatory desynchronisation is known to span domains larger than the phoneme, it is more adequate to employ higher-level units such as syllables in order to combine features. The temporal alignment of features within syllables or comparable units is highly variable; for this reason, overlap relations between features are left completely unspecified ("desynchronised") in our system. Syllables are thus defined as parallel sequences of features, with the sole requirement that these sequences begin and end simultaneously. These syllable templates are then mapped to feature-coded entries in the recognition lexicon by means of a multi-level string-matching algorithm.

2. Design of the Feature-Based Recogniser

2.1. Phonetic Features

The system developed consists of a number of serial modules: a feature-recognition front end, a synchronisation module and a lexical mapping component. The phonetic features shown in Table 1 serve as basic recognition units. These features are contrastive in German, i.e. they serve to distinguish lexical items; however, non-contrastive, allophonic features may in principle be included as well. For each feature-value a separate hidden Markov model (HMM) is trained; in addition to this, a silence model (sil) is employed.

Feature	**Feature-values**
phonation	voiced, voiceless
manner	occlusive, fricative, lateral, nasal, vowel
place	labial, coronal, palatal, uvular, glottal, high, mid, low
front-back	front, back, nil
roundness	rounded, unrounded, nil
centrality	central, non-central, nil

Table 1: Phonetic features for German

Feature-values are grouped into six classes defined by the phonetic features which subsume them: *phonation*, *manner*, *front-back*, *roundness* and *centrality*. Within each class, feature-value HMMs are employed disjunctively during training and recognition, i.e. for each signal frame a decision is enforced in favour of one feature-value to the exclusion of all others. Across classes, however, HMMs are arranged in parallel: HMMs belonging to different classes are executed simultaneously. Thus, the feature detection front end outputs six parallel sequences of feature-values (one for each class).

2.2. (De)synchronisation and Lexical Access

Feature-value sequences are enriched with information about syllable boundaries. Together, this information is used in order to synchronise features within identical syllable boundaries. Syllables are thus defined as parallel stretches of feature-value sequences, similar to descriptions in non-linear phonology (c.f. e.g. [5, 6]). However, no attempt is made to characterise the precise temporal alignment

of individual feature-values. The result of feature synchronisation is a sequence of temporally underspecified syllable templates. The reason for choosing syllables rather than phonemes or words as synchronisation units is that they cover more coarticulatory variation than phonemes. On the other hand, they form a finite set, as opposed to words, which form a potentially infinite set in any language.

In a second step these templates are passed on to a lexical access module which maps them to entries in the syllable recognition lexicon using a multi-level dynamic programming algorithm. The recognition lexicon consists of syllable entries which are equally coded as six parallel sequences of features. For each of these sequences, the edit distance to the corresponding sequence in the syllable hypothesis template is computed using dynamic programming. Individual distance values are then summed up to an overall distance value per entry, with the possibility of selectively weighting certain sequences, i.e. phonetic classes. The N entries with the lowest distance values are then selected for further processing, i.e. for evaluation or (in the future) for word and sentence recognition.

The feature-based lexicon offers the advantage of being able to process speech variants easily. Many variants are distinguished by the presence vs. absence of relatively few feature-values; these can be accommodated in the lexical representation of items by specifying optional or disjunctive feature-values. The disadvantages of introducing optional or disjunctive values are the increased complexity of the lexical mapping algorithm as well as possible confusions between lexical items. In order to maintain a good performance of the system the inclusion of optional and disjunctive elements must be carefully monitored.

3. Data and Implementation

The system was tested on a corpus of spontaneous speech (German) produced by eight male and two female speakers. The data consisted of scheduling dialogues between two interlocutors recorded within the context of Verbmobil automatic translation project [8].

The size of the training material was 16 hrs; the size of the test set was 1 h 30 mins. Preprocessing, feature training and feature recognition were carried out by a commercially available HMM toolkit ("HTK", [11]). The data was sampled with 16 kHz and low-pass filtered at 8 kHz. Twenty-four mel-frequency cepstral coefficients were extracted at a frame rate of 10 ms using a 16 ms Hamming window. First-order differentials and an an energy component were used. Feature HMMs were implemented as left-to-right models with three to five states. Output probabilities were modelled by single Gaussian probability density functions (PDFs). Initialisation was carried out using feature labels derived from manually-produced phoneme labels. No language model was used during feature recognition; however, a linguistic recognition network defining permissible feature sequences was employed.

The lexical mapping component accesses a syllable recognition lexicon consisting of 600 canonical entries, enriched with optional and disjunctive feature specifications to accommodate speech variants. These cover the most frequent fluent speech phenomena to be found in the speech corpus: glottal stop elision, reduction of final schwa syllables, reduction of function words, vowel reduction and elision of final coronal stops.

A triphone-based HMM recogniser was trained and tested on the same data set, using identical preprocessing parameters. 56 three-state left-to-right phoneme models were initialised and trained on hand-labelled data. These were then cloned to yield 3382 triphone models which subsequently underwent re-estimation. Output distributions were approximated by five mixture components. During phoneme recognition a bigram model was employed.

4. Results

Detailed feature recognition results are listed in Table 2. The average feature recognition rate is 91.8%; recognition rates are best for phonation features whereas place features are the least robust.

Phonation					
+voi	98.66	-voi	100.00		
Centrality					
+cent	71.22	-cent	89.71	nil	96.40
Roundness					
+rnd	93.22	-rnd	80.21	nil	93.85
Front-Back					
front	87.69	back	95.90	nil	97.33
Manner					
fric	91.30	occ	93.44	nas	92.55
lat	83.33	vo	93.67	sil	97.96
Place					
cor	87.93	glott	90.62	high	89.87
lab	91.84	low	97.94	mid	88.62
pal	100.00	uvu	93.94	vel	88.89

Table 2: Feature recognition rates

The syllable recognition rates for the feature-based recogniser are shown in Table 3, as well as the recognition rates for the phoneme sequence derived from the top syllable sequence. (Table 3).

The results obtained by the triphone-based recogniser are given in Table 4.

	Correctness	Accuracy
Syllables	48.1%	48.1%
Phonemes	73.7%	68.3%

Table 3: Phoneme and syllable recognition rates – feature-based recogniser

	Correctness	Accuracy
Phonemes	64.84 %	54.81 %

Table 4: Phoneme rates – triphone-based recogniser

5. Summary

The system presented relies on phonetic features as basic recognition units and combines these at the syllable-level to form temporally underspecified syllable templates. These are then mapped to entries in a feature-based syllable recognition lexicon using multi-level dynamic programming. The advantage of late synchronisation of features resides in the better modelling of coarticulation, which is caused primarily by temporal misalignments of articulatory movements. Tests on a medium-sized corpus of spontaneous speech (German) in comparison with a triphone-based recogniser revealed a superior performance of the feature-based recogniser for the present data set. It should be pointed out that the feature-based recogniser makes use of a very simple kind of statistical modelling, viz. single Gaussian PDFs, and does not employ a language model, whereas the triphone-based recogniser does use mixture densities and statistical a priori constraints in the form of a bigram model. We may conclude from this that coarticulatory modelling is more effectively carried out in the feature-based recogniser. Moreover, only small number of feature models is required compared to a large number of triphone models. Future tests will have to show whether the feature-based approach is applicable to very large vocabularies. Further extensions to the present system will include a prosody component and the development of an incremental architecture, allowing the exchange of feedback between different modules.

6. References

1. P. Dalsgaard. Phoneme label alignment using acoustic-phonetic features and Gaussian probability density functions. *Computer, Speech and Language 6*, pages 303–329, 1991.
2. P. Dalsgaard, O. Andersen, and W. Barry. Multilingual label alignment using acoustic-phonetic features derived by neural-network technique. *Proceedings ICASSP-92*, pages 197–200, 1992.
3. L. Deng and K. Erler. Hidden markov model representation of quantized articulatory features for speech recognition. *Computer, Speech and Language 7*, pages 265–282, 1993.
4. L. Deng and K. Erler. Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds. *Proceedings ICASSP-94*, pages 45–48, 1994.
5. J. Goldsmith. *Autosegmental Phonology*. Garland Press, New York, 1979.
6. J. Goldsmith. *Autosegmental and Metrical Phonology*. Blackwell, Oxford, 1990.
7. K. Hübener and J. Carson-Berndsen. Phoneme recognition using acoustic events. *Proceedings ICSLP-94*, pages 1919–1922, 1994.
8. K. Kohler, G. Lex, M. Pätzold, A. Simpson, and W. Thon. *Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL – 3.0*. IPDS Kiel, VM Technisches Dokument Nr.11, 1994.
9. O.Schmidbauer. Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations. *Proceedings ICASSP-89*, pages 616–619, 1989.
10. G. Pérennou, H. Kabre, and N. Vigouroux. *Automatic SAPHO system segmentation of EUROM-0 multilingual speech corpora into phonetic events*. Internal Report III.b of ESPRIT PROJECT 2589 (SAM), 1991.
11. P.C. Woodland S.J. Young and W.J. Byrne. *HTK Version 1.5: User, Reference and Programmer Manual*. Publ. Entropic Research Laboratories, Washington DC, 1993.