FEATURE SELECTION WITH EXCEPTION HANDLING — AN EXAMPLE FROM PHONOLOGY

GABRIELE SCHELER
Forschungsgruppe KI/Kognition, Institut für Informatik
Technische Universität München
Waldeslust 4, 81377 München, Germany
e-mail: scheler@informatik.tu-muenchen.de

ABSTRACT

The goal in this paper is to show how the classification of patterns of phonetic features (=phones) to phonemes can be acquired. This classificational process is modeled by a supervised feature selection method, based on adaptive distance measures. Exception handling is incorporated into a learned distance function by pointwise additions of Boolean functions for individual pattern combinations. An important result is the differentiation of **rules** and **exceptions** during learning.

1 PHONETIC FEATURES AND PHONEMES

The goal in this paper is to show how the classification of patterns of phonetic features (=phones) to phonemes can be acquired. In every language a number of differentiable phones belong to one phoneme. To learn the phonemic pattern of a language amounts to learn a classification of all naturally occurring phones to a phoneme.

The continuum of articulatory places or acoustically defined frequency formants for a single phone can be cut up into a set of descriptive features ([1]). The phonetic representation chosen in this example is a rather conservative one, based on the IPA-notation. Similar results should be obtainable with other phonetic representations, derived more directly from acoustic speech analysis. At least motor representations usually contain all the features that are necessary to distinguish phoneme classes.

Phonetic features for German vowels, the corresponding patterns and their intended classification are shown in Table 1, adapted from [5]. Only tense vowels are shown. Non-segmental features such as length, syllabicity, nasalization, rhotacization, diphthongs have also been excluded. Note that increasing the number of features does not affect the analysis, as further distinctions will simply be ignored during learning of the classification.

As has been noted by phonologists for a long time ([10]), the grouping of phones to phonemes gives rise to phonological systems, which are organized by few phonetic

Table 1: Phonetic feature descriptions for allophonic variations of German vowels

	front rounded	front unrounded	central unrounded	back unrounded	back rounded
high	ue	i			u
mid-high	oe1	e1	e2		o1
mid-low	oe2	ae1		a2	o2
low		ae2	a1	a3	o3

contrasts, and consist of phonemes with more or less systematicity.

This classificational process is modeled here by a supervised feature selection method, i.e. classification by a subset of the original features which is based on a training set of classified patterns.

In general, feature selection means to find a set of relevant features for classification in contrast to feature extraction where features are extracted as linear combinations of existing features (cf. [4],pp. 246–248, [2],pp. 187–190). I.e. in this case features are not constructed by the phonological analysis component (which is cognition rather than perception) after the phonetic representation has been established.

We are using a parametrized distance function as an adaptive measure of similarity and a corresponding training scheme to set its parameters in order to synthesize a specific distance function (cf. [9] for a detailed analysis of the learning scheme, cf. also [8]).

2 CLASSIFICATION OF GERMAN VOWELS

Input patterns have been encoded as binary feature vectors of length 9, where each position marks the absence or presence of a phonetic feature from Table 1.

E.g., “i”

Front	Central	Back	Rounded	Unrounded	High	Mid-High	Mid-Low	Low
1	0	0	0	1	1	0	0	0

The distance function scheme for feature selection is a weighted Hamming distance called d_λ and defined as:

$$d_\lambda(pattern1, pattern2) = \sum_i^n \lambda_i \times |pattern1_i - pattern2_i|$$

For each of the target classes a single exemplar is given in order to achieve a nearest-neighbor classification with the learned distance function for all other patterns.

Learning (or “adapting”) a distance function is achieved by setting values to the parameters λ_i.

All λ_i that occur (as a sum) in the distance equation for patterns of the same class are set to 0, and an arbitrary free variable λ_i of those occurring in the sum for patterns of different classes is set to 1.

This very simple learning method has certain advantages:

1. λ_i corresponding to irrelevant features are set to 0 or left unset. Accordingly the size of the feature set does not affect learning or generalization.

2. For a certain unknown parameter λ_i we can define a training pattern x that gives a value to it, by changing the i'th position in a classified pattern p: $d_\lambda(x,p) = \lambda_i$.

3. By monitoring conflicts in parameters settings, we can make a list of patterns not representable by the distance function scheme.

This is useful because we can make a list of exceptions while learning, and we notice when patterns to set a parameter are not within the problem space. Minimal training sets can be generated from the classified patterns(cf. 2). Here we used a training set consisting of 3 patterns (oe1, a2, a3).

The values for the parameters after training are :

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9
1	0	0	1	λ_5	1	0	0	0

There are no patterns available to set λ_5.

In applying the distance function we find that patterns for the phoneme 'e' (e1,e2) are not correctly classified, all others are.

These results are closely related to equivalent terms of classical structural phonology:

- Certain phonetic contrasts are "neutralized", namely center/back, mid-high/ mid-low/low. The phonological contrasts are front vs. (center, back) and high vs. (mid-high, mid-low, low).

- Rounded/Unrounded form a 'privative' pair, i.e. it is sufficient to mark one of them.

- The phoneme 'e' has less systematicity, therefore secondary phonological contrasts have to be introduced to characterize it in addition to the fully systematical vowel phonemes.

The result of the learning process can also easily be transformed into a classical phoneme definition operating with distinctive (phonological) features.

Distinctive features of the German vowel system are according to this analysis, (cf. also [6], [11])

$$\begin{bmatrix} +-front \\ +-rounded \\ +-high \end{bmatrix}$$

When we apply a learning procedure, in some cases we are content with a certain percentage of correctness. In this case, 13 of 15 patterns are correctly classified, which is 86,7%. But at least in the context of linguistic abilities we need a method which allows to acquire a certain faculty completely.

3 EXCEPTION HANDLING

By using d_λ as the basis for classification, we have imposed a specific classification and generalization scheme. I.e. only certain classifications can be learned with λ-parameters, and the generalization principle from training examples to a whole pattern set is also fixed.

When a training set is not fully representable with λ-parameters, there is the option of representing conflicting training patterns explicitly and separately as "exceptions" within the overall distance function. In comparison to the option of choosing a distance function scheme with higher representational power (a more general function approximator) this has the advantage that learning effort can be kept low, rule abstraction is still simple, and generalization restricted to the "regular patterns" rather than the examples.

In a situation, where all exceptions can be listed as training examples, 100% correctness of the learned pattern classifier can be achieved.

There are two cases of misclassifications of two patterns a and b without exception handling:

- $d_\lambda(a, b) = 1$ where it should be $= 0$.

  The goal is to multiply the distance function by a term, which yields 0 when a and b are compared and 1 otherwise.

  The simplest expression for the multiplication term is the Boolean function $f_{\neg(ab)}$ which is $= 0$ iff $x =< ab >$ and which can be written as

  $$f_{\neg(ab)}(x) = \neg(\bigwedge_i^n \chi_i \wedge \bigwedge_i^n \epsilon_i)$$

  $$\begin{array}{llll} \chi_i = x_i & \text{if } a_i = 1 & \epsilon_i = y_i & \text{if } b_i = 1 \\ \chi_i = \neg x_i & \text{if } a_i = 0 & \epsilon_i = \neg y_i & \text{if } b_i = 0 \end{array}$$

  By substituting $\neg x$ by $(1 - x)$ and $\wedge$ by $\times$ we can translate this function into an arithmetic expression. The Boolean function $f_{\neg(e1e2)}$ requires $2n$ (n = patternlength, here: 18) variables.

- $d_\lambda(a, b) = 0$ where it should be $= 1$.

  Here we want to add a term to the distance function which is 1 for a and b as input and 0 otherwise. Again the simplest expression is a function f_{ab}, defined like $f_{\neg(ab)}$ above.

For the given problem, this would require to add the functions (f_{ae1e1}, f_{ae2e1}, f_{a1e2}, f_{a2e2}, f_{a3e2}) with a total of 5*18 variables.

Result:

$$d_\lambda^{excep}(x,y) = d_\lambda(x,y) \times \prod_{\forall a,b} f_{\neg(ab)}(< xy >) + \sum_{\forall a,b} f_{ab}(< xy >)$$

There are several possibilities of simplifying this expression (multiplicational as well as additional terms) based on the following observation: Each individual Boolean function f may be represented by another function f^* as long as the following equations hold:

$f^*_{ab}(< xy >) = 1 = f_{ab}(< xy >)$ for patterns $a = x, b = y$

$f^*_{ab}(< xy >) = 0 = f_{ab}(< xy >)$ for patterns x, y such that $class(x) = class(y)$

$f^*_{ab}(< xy >)$ = arbitrary value, for all other patterns ($class(x) \neq class(y)$ or x, y do not occur in problem space)

In a similar vein we can define $f^*_{\neg(xy)}$. From this observation it follows that simplifications (i.e. using fewer variables in the definition of the Boolean function) have to be performed dependent on the actual classes that exist.

The following strategy has been used for the implemented function definition:

1. find a position i where a and b differ.
2. for all patterns x, $class(x) = class(a)$: are all x equal at position i? If not, go to 1.
3. for all patterns y, $class(y) = class(b)$: are all x equal at position i? If not, go to 1.
4. position i found.

The strategy fails, when all positions have been tried and step 4 is not reached. Otherwise a positional differential criterion i for the exceptional patterns has been found. Then we can define an additional term for $d(a,b)$ as:

$$+|x_i - y_i| * desc(a)$$

$$desc(a) = \neg x_1 \wedge \neg x_4 \wedge \neg x_6$$

$desc(a)$ is a description for exactly all patterns in $class(a)$, which is given by the result of the learning of d_λ.

Simplifications of Boolean functions allow in principle the introduction of 'secondary rules', i.e. a generalization from the exceptions. Because such a simplified function is not maximally constrained, it incorporates by itself a certain limited generalization. For instance, the new positional criterion i can be the basis for a secondary rule schema.

This possibility has not been further pursued here.

The complete distance function for the classification of phonetic feature vectors to German vowel phonemes, gained by this method is:

$$\begin{aligned} d_\lambda^{excep}(x,y) = \; & \sum_i^n \lambda_i * |x_i - y_i| * \\ & 1 - (x_1(1-x_2)(1-x_3)(1-x_4)x_5(1-x_6)x_7(1-x_8)(1-x_9) * \\ & (1-y_1)y_2(1-y_3)(1-y_4)y_5(1-y_6)y_7(1-y_8)(1-y_9)) + \\ & x_1(1-x_4)(1-x_6)|x_7-y_7| + (1-x_1)(1-x_4)(1-x_6)|x_7-y_7| \end{aligned}$$

4 RESULTS OF ANALYSIS

In the processing of speech, acoustic input is transformed into a phonetic representation, consisting of individual phonetic features. This process is also known as categorical perception. The task of phonology consists of relating these universal phonetic representations to the language–specific phonological units, e.g. phonemes, which operate as true symbols in the language process.

In particular, we use the observation made in structural phonology that few phonetic contrasts are actually needed to determine the phonemic class. A classification using a feature selection technique shows which phonetic features are employed as phonological contrasts.

The fact that we can automatically reproduce the results of classical structural phonology points to the idea of applying this method to lexical semantics as well. The view of the lexicon as a set of concepts consisting of feature sets, employing distinctive features for lexico-syntactic classes etc. might be put to an empirical test.

With the indicated encoding and the distance function scheme d_λ we achieve a separation as required. However, certain patterns may not be classifiable with this type of generalization. Rather than extending the feature selection model it seems justified to treat these cases as "exceptions", which are a common and real occurrence in linguistic functions.

After parameters in the distance function scheme have been set, exceptional cases are defined and can be recognized by the classifier.

The distance classifier can then be automatically enhanced to cover all the exceptional cases that occur.

In contrast to prevailing methods of learning from examples (e.g. function approximation by back-propagation [7] or machine learning techniques [3]), this method provides a way of distinguishing between generalizable features which form the basis of the distance function scheme and additionally stored patterns, which are added as deviating function values. Thus one possibility of distinguishing rules from exceptions within a pattern classification framework is presented.

REFERENCES

[1] Barry,W.J. and A.J. Fourcin: Levels of labelling. *Computer Speech and Language* **6**(1992) :1-14.

[2] Devijfer,P.A. and J.Kittler: *Pattern Recognition. A statistical approach.* Prentice Hall 1982.

[3] Dietterich,T. and R. Michalski: A Comparative Review of selected Methods for Learning from Examples. In: Michalski,R., J.Carbonell and T. Mitchell: *Machine Learning. An Artificial Intelligence Approach.* Springer 1984.

[4] Duda,R.O. and P.E. Hart: *Pattern Classification and Scene Analysis.* John Wiley 1973.

[5] Hyman,L.: *Phonology. Theory and Analysis.*Holt, Rinehart and Winston 1975.

[6] Kohler,K.: *Phonetik des Deutschen.* 1977.

[7] Rumelhart,D.,G.Hinton and R. Williams: Learning Internal Representations by Error Propagation. In: McClelland,J. and D.Rumelhart: *Parallel Distributed Processing.* MIT Press 1986.

[8] Scheler,G.: The Use of an Adaptive Distance Measure in Generalizing Pattern Learning. *Proceedings of ICANN92*, Elsevier 1992,pp. 131-134.

[9] Scheler,G.: *Pattern Classification with Adaptive Distance Measures.* forthcoming 1993.

[10] Trubetzkoy,N.S.: *Grundzüge der Phonologie.* Vandenhoeck und Ruprecht 1958, 6th ed. 1977.

[11] Waengler,H.-H.: *Grundriss einer Phonetik des Deutschen.* Marburg, 4th ed. 1983.