# Exploring the connection of acoustic and distinctive features

*Thomas Kisler, Uwe D. Reichel*

Institute of Phonetics and Speech Processing, Munich University, Munich, Germany

{kisler,reichelu}@phonetik.uni-muenchen.de

## Abstract

This study is a contribution to link the abstract phonological level to the acoustic signal level by identifying the main acoustic correlates for the distinctive feature set developed by Chomsky and Halle (1968). The acoustic features were extracted by the openSMILE toolkit from spontaneous speech data. For each distinctive feature a set of closely related acoustic features was derived by means of correlation-based feature selection. Based on the respective acoustic feature pools C4.5 trees and support vector machines for binary feature classification were trained. The classification performance ranged from 76 to 89% for vocalic features and from 78 to 93% for consonantal features. The methods proposed in this study can be of use to identify systematic speech signal correspondencies for phonological models and as a starting point for distinctive feature detection in speech recognition.

**Index Terms**: distinctive features, acoustic features, feature selection, machine learning

## 1. Introduction

On the phonological level phonemes can be characterized by sets of distinctive features. These features had been defined in acoustic terms by Jakobson [1] or in articulatory terms by Chomsky and Halle [2].

Since these high-level distinctive features (DF) are an abstract representation of the phoneme inventory of a language, their relation to the acoustic speech signal is not straightforward. Attempts to shed light on this relation can be divided into an expert- and a data-driven group. Expert-driven approaches as [1, 3] rely on acoustic phonetic knowledge to carefully choose a small number of complex acoustic features (AF) like the energy in relevant spectral regions and acoustic discontinuities.

Data-driven approaches in contrast make use of machine learning methods like neural networks [4] to predict DF from a larger number of low-level AFs.

DF detectors are applied in phoneme and speech recognition. [5] used a Kohonen net to map AFs to DFs that in turn were used to train a HMM for phoneme recognition. In [6] acoustics was mapped to DF by means of Multi-Layer perceptrons providing the input for Time-Delay Neural Networks for phoneme recognition. [7] extracted a discriminative DF subset to train HMMs for speech recognition. [8] showed that the usage of DF can turn speech recognition systems to be more robust under noisy conditions.

We pursue a data-driven, exploratory bottom-up approach to find mappings from acoustics to DF of Chomsky and Halle's feature system, which still is widely-used for sound system descriptions. The approach is exploratory in such a way that we start from a very large number of AFs (106 in total) among which we infer the most relevant ones for each DF by means of a feature selection method. Thereby for each DF its major acoustic correlates were extracted, which were subsequently taken as input for decision tree and support vector machine DF classifiers. To additionally explore the power of non-standard features, we also took the musical chroma features (and their delta and delta delta regression), into account. The study is designed to explore the connection between AFs and DFs which could be used by phoneticians and engineers that are working on improving the quality of automated systems, like phoneme or speech recognition.

## 2. Data

As the basis for data extraction the Kiel Corpus of spontaneous speech [9, 10] was used. The corpus consists of 52 German speakers resulting in roughly 2000 dialog turns, in which two speakers have to complete a scheduling task.

The corpus is manually segmented and labeled and contains well over 200,000 phonemes which we assumed being enough for the evaluation of the proposed method.

## 3. Method

### 3.1. Overview

To estimate which AFs seem to be most prevalent when a certain DF is activated, for every DF and its associated phonemes, we extracted a set of AFs from the signal, used a ranking to select the $n$ best features that are able to predict the respective class and tested the discriminative power of the resulting AF subset with a classificator.

### 3.2. Distinctive features

To test our method, we examined the connection between DFs and AFs for the group of a) the sonorants, glides and laterals and b) for vowels depicted in table 1 and 2. For the remainder of this paper whenever we speak of consonants, we mean the evaluated sonorants, glides and laterals. Based on the previously mentioned models, the features are either binary or privative. A binary features has two possible values "+" and "-". Privative features are either present or absent, and are signaled by check signs and capital abbreviations. A complete overview and a description of the features can be found in [2, 11].

We did not include the feature classes *lat*, *appr* and *DORS* as the classes were to unbalanced regarding their occurrence in the data, which otherwise would have resulted in an over fitting of the machine learning algorithms and misclassification of a complete class. Furthermore, the DFs *voiced* and *asp* are inverse to each other and therefore produce the same results.

### 3.3. Acoustic features

For the AF extraction we used openSMILE, an open-source tool for feature extraction from audio signals developed at the

Table 1: *Distinctive features for the German sonorants, glides and laterals for which we examined the connection to AFs.* cons *means consonantal,* appr *approximant,* cont *continuant,* nas *nasal,* LAB *labial and* COR *coronal*

| P | *cons* | *voiced* | *cont* | *nas* | *LAB* | *COR* |
|---|---|---|---|---|---|---|
| m | + | + | - | + | ✓ | |
| n | + | + | - | + | | ✓ |
| ŋ | + | + | - | + | | |
| l | + | + | - | - | | ✓ |
| r | + | + | + | - | | ✓ |
| j | - | + | + | - | | ✓ |
| h | - | - | + | - | | |

Table 2: *Distinctive features for German vowels for which we examined the connection to AFs.* LAB *means labial.*

| P | *back* | *high* | *low* | *tense* | *LAB* |
|---|---|---|---|---|---|
| i | - | + | - | + | |
| ɪ | - | + | - | - | |
| y | - | + | - | + | ✓ |
| ʏ | - | + | - | - | ✓ |
| e | - | - | - | + | |
| ɛ | - | - | - | - | |
| ø | - | - | - | + | ✓ |
| œ | - | - | - | - | ✓ |
| æ | - | - | + | - | |
| u | + | + | - | + | ✓ |
| ʊ | + | + | - | - | ✓ |
| o | + | - | - | + | ✓ |
| ɔ | + | - | - | - | ✓ |
| a | + | - | + | - | |
| ɑ | + | - | + | - | |

Technische Universtät München. It provides a big set of readily available AFs that can be extracted from speech data [12].

We used the following atomic and complex AFs to be evaluated for their power in describing certain DFs.

- *RMSenergy:* Root-mean-square energy
- *LOGenergy:* Logarithmic energy
- *ZCR:* Zero-crossing rate
- *MCR:* Mean-crossing rate
- *Intensity:* Simplified frame intensity
- *F0:* Fundamental frequency
- *HNR:* Harmonics-to-noise ratio
- *voiceProb:* Probability of voicing
- $mfcc_{0-12}$*:* Mel-frequency cepstral coefficients (coeff.)
- $mfcc\Delta_{0-12}$*:* Delta regression from MFCC
- $mfcc\Delta\Delta_{0-12}$*:* Delta delta regression from MFCC
- $chroma_{0-11}$*:* CHROMA features from semi-tone scaled spectrum
- $chroma\Delta_{0-11}$*:* Delta regression of CHROMA features
- $chroma\Delta\Delta_{0-11}$*:* Delta delta regression of CHROMA features
- $lpcCoeff_{0-7}$*:* Linear predictive coding coeff.
- $FFLpc_{0-6}$*:* Formant frequencies and bandwidth from LPC coeff.
- $lspFreq_{0-7}$*:* Line spectral pairs (LSP) from LPC coeff.

These features left us with 106 different AF values that are computed directly based on the signal or derived from other AFs. For a detailed description of the available features in OpenSMILE and an in-depth description of the ones used can be found in [13].

We extracted the AFs within $25ms$ windows located around the segments' temporal center. There the closest match to the articulatory targets and the least degree of co-articulation is to be expected which gives the closest-possible correspondence to the segments' phonological representation as shown in table 1 and 2. Note, that in this initial attempt relevant transitional acoustic cues at segment boundaries as well as segments with highly non-static acoustic characteristics (as plosives and diphthongs) were left aside. These issues will be addressed in follow-up studies.

### 3.4. Acoustic feature selection

As execution time was a crucial point for the choice of the feature selection algorithm, we decided to use a simple method based on correlation (Pearson product), implemented in the Waikato Environment for Knowledge Analysis (WEKA) toolkit [14]. The ranking is calculated for each DF on the correlation coefficient between the dichotomous class and the actual feature values, where higher correlation results in higher ranking.

The robustness of feature selection was successfully tested by running an alternative selection algorithm based on the OneR classifier explained in [15]. A comparison of both selection methods has shown, that they only differ marginally in both the selected features and therefore in the resulting classification results. The mean difference of the classification accuracy on AFs selected by correlation and OneR was $0.46\%$.

We conducted three runs of our evaluation, where we subsequently selected the $n$ best features for $n \in \{3, 15, 106\}$.

### 3.5. Classification with a reduced acoustic feature set

After the selection of the $n$ best features for each DF, we evaluated the resulting AF set with regard to its performance in a two-class classification test. For binary features those were "0" and "1", for privative features "on" and "off". To eliminate the chance of biasing the result by the choice of the classificator we used both a C4.5 tree [16] and a Support Vector Machine (SVM) [17] with a first order polynomial kernel.

Based on the default settings for WEKA the SVM normalized the training data during learning. For both algorithms we executed a 10-fold stratified cross-validation for each DF.

## 4. Results

### 4.1. DFs of the selected consonants

The results for the evaluated consonants as described in table 1, are shown in table 3. The classification results of the Support Vector Machine and the C4.5 tree are very similar. The C4.5 tree shows better results in the feature subset whereas the SVM shows better results when the full feature set is evaluated. They range for 3 features between 87.76 and 89.58%, between 91.83 and 92.56% for 15 features, and between 91.47 and 94.13% for all features. The best 3 features set consists of the same features, only the position in the ranking varies between the top two features. The results are as expected and improve for more features, but the improvement is smaller when comparing 15 and 106 features, than it is from 3 to 15 features.

The top AFs shared between all classes (by ranking order)

Table 3: *Results of evaluation of binary features on sonorants, glides and laryngals on the 3 and 15 best and all features. The class ratio is DF class '+' divided by phoneme class '-'. To save space, in the rows where the 15 best features are shown, the first three are ommited, as they are the same as in the column of the 3 best features.*

| Distinctive features | *cons* | *voiced* | *cont* | *nas* |
|---|---|---|---|---|
| Phoneme class '+' | [r, n, l, ŋ, m] | [r, n, l, ŋ, m, j] | [r, j, h] | [n, ŋ, m] |
| Phoneme class '-' | [j, h] | [h] | [n, l, ŋ, m] | [r, l, j, h] |
| Class ratio | 1.419 | 1.626 | 0.831 | 0.895 |
| Top 3 Features | $mfcc_2$ | $lspFreq_2$ | $mfcc_2$ | $lspFreq_2$ |
| | $lspFreq_2$ | $mfcc_2$ | $lspFreq_2$ | $mfcc_2$ |
| | HNR | HNR | HNR | HNR |
| Top 3 Class. C4.5 | 88.58% | 88.99% | 89.58% | 88.35% |
| Top 3 Class. SVM | 88.19% | 88.70% | 89.38% | 87.76% |
| Top 15 Features | [...] | [...] | [...] | [...] |
| | $mfcc\Delta\Delta_2$ | $mfcc\Delta\Delta_2$ | $mfcc\Delta\Delta_2$ | MCR |
| | MCR | MCR | MCR | voiceProb |
| | voiceProb | voiceProb | voiceProb | $lspFreq_1$ |
| | $mfcc\Delta_1$ | $lspFreq_1$ | $lspFreq_1$ | $mfcc\Delta\Delta_2$ |
| | $lspFreq_1$ | $mfcc\Delta_1$ | $lspFreq_4$ | $mfcc_{10}$ |
| | $lspFreq_4$ | $lspFreq_4$ | $mfcc\Delta_1$ | $lspFreq_4$ |
| | $mfcc\Delta\Delta_1$ | $mfcc\Delta\Delta_1$ | $mfcc_{10}$ | $mfcc_9$ |
| | $lpcCoeff_1$ | $lpcCoeff_1$ | $lpcCoeff_1$ | $chroma\Delta_0$ |
| | $lspFreq_3$ | $chroma\Delta\Delta_{10}$ | $lspFreq_3$ | $chroma_{10}$ |
| | $chroma\Delta\Delta_{10}$ | $mfcc_{10}$ | $mfcc_9$ | $mfcc\Delta_1$ |
| | $chroma_{10}$ | $lpcCoeff_0$ | $chroma_{10}$ | $chroma\Delta\Delta_{10}$ |
| | $chroma\Delta_0$ | $lspFreq_3$ | $chroma\Delta_0$ | $chroma_9$ |
| Top 15 Class. C4.5 | 92.07% | 92.56% | 92.34% | 91.91% |
| Top 15 Class. SVM | 91.83% | 92.21% | 92.20% | 92.35% |
| Features | all | all | all | all |
| All Class. C4.5 | 91.80% | 92.30% | 91.84% | 91.47% |
| All Class. SVM | 93.07% | 93.37% | 94.02% | 94.13% |

are second MFCC, second LSP, harmonics-to-noise ratio, delta delta regression of second MFCC ($mfcc\Delta\Delta$), MCR, voicing probability, delta regression of first MFCC ($mfcc\Delta$), first and fourth LSP can be found within all DFs, with different ranking.

The results of the evaluation of the privative DFs *LAB* and *COR* can be found in table 4. The correct classification ranges between 77.8 and 86.68%. The feature selected AFs are not very stable between the two DF classes. The differences between C4.5 and SVM are only marginal, yet the decision tree performs slightly better.

### 4.2. DFs of vowels

The results of the evaluation of binary DFs for German vowels as described in table 2, are shown in table 5. The classification results range from 76.46 to 88.6%. The top AFs are not very stable when regarding the different DFs. The classification results again do not differ much, though the SVM performs slightly better in vowels than the C4.5 tree. Opposed to the consonant results, the classification performance for the only available privative feature is as good as for most of the binary features.

## 5. Discussion

### 5.1. DFs of the selected consonants

The top 3 AFs for the sonorants, glides and laterals are the second MFCC coefficient, the second LSP and the harmonics-to-noise ratio (HNR). Interestingly enough, that the best 3 AFs for all DFs are the same. This means that in general those AFs seem to have the most descriptive power regarding DFs for the

Table 4: *Results of evaluation of privative features on sonorants, glides and laryngals based on the 15 best features. The class ratio is phoneme class '✓' divided by phoneme class ' '. Phon. is short for phoneme, Class. short for Classificator and Dist. features for distinctive features.*

| Dist. features | *LAB* | *COR* |
|---|---|---|
| Phon. class '✓' | [m] | [r, n, l, j] |
| Phon. class ' ' | [r, n, l, ŋ, j, h] | [ŋ, m, h] |
| Class ratio | 0.155 | 0.847 |
| Features | $mfcc_2$ | $mfcc\Delta\Delta_2$ |
| | $lspFreq_2$ | $lspFreq_2$ |
| | $mfcc_{10}$ | $mfcc\Delta_1$ |
| | HNR | $mfcc_2$ |
| | $lspFreq_1$ | HNR |
| | MCR | MCR |
| | voiceProb | voiceProb |
| | $lspFreq_4$ | $mfcc\Delta\Delta_1$ |
| | $mfcc_5$ | $lspFreq_1$ |
| | $lspFreq_3$ | $lspFreq_4$ |
| | $mfcc\Delta_0$ | $lpcCoeff_1$ |
| | $chroma_{10}$ | $chroma\Delta\Delta_{10}$ |
| | $chroma_9$ | $chroma\Delta\Delta_3$ |
| | $mfcc_9$ | $lpcCoeff_0$ |
| | $mfcc_0$ | $mfcc_6$ |
| Class. C4.5 | 86.85% | 78.23% |
| Class. SVM | 86.68% | 77.80% |

Table 5: *Results of evaluation of binary features on vowels based on the 15 best features. The class ratio is phoneme class + divided by phoneme class -. Phon. is short for phoneme, Class. short for Classificator and Dist. features for distinctive features.*

| Dist. features | *back* | *high* | *low* | *tense* | *LAB* |
|---|---|---|---|---|---|
| Phon. class + | [a, o, ɔ, ʊ, ɑ, u] | [i, ɪ, ʊ, y, ʏ, u] | [a, æ, ɑ] | [ø, i, e, o, y, u] | [ø, o, ɔ, ʊ, y, œ, ʏ, u] |
| Phon. class - | [ɛ, i, e, ɪ, æ, y, œ, ʏ] | [ɛ, a, e, o, ɔ, æ, ɑ, œ] | [ɛ, i, e, o, ɔ, ɪ, ʊ, y, œ, ʏ, u] | [a, ɔ, ɪ, æ, ʊ, ɑ, œ, ʏ] | [a, i, e, ɪ, æ, ɑ] |
| Class ratio | 1.135 | 0.577 | 0.55 | 0.336 | 0.327 |
| | $lspFreq_3$ | $lspFreq_1$ | $lspFreq_1$ | $lspFreq_1$ | $lspFreq_3$ |
| | $mfcc_4$ | $mfcc_4$ | $mfcc_4$ | $mfcc_4$ | $mfcc_2$ |
| | $mfcc_6$ | $mfcc_3$ | $mfcc_6$ | $mfcc_3$ | $lspFreq_2$ |
| | $lspFreq_1$ | $lspFreq_3$ | $mfcc_3$ | $mfcc\Delta\Delta_4$ | $lspFreq_0$ |
| | $mfcc\Delta\Delta_4$ | $mfcc\Delta\Delta_4$ | $lspFreq_3$ | $mfcc_6$ | $FFLpc_1$ |
| | $lspFreq_0$ | $chroma_8$ | $chroma_8$ | $lspFreq_3$ | $mfcc_4$ |
| | $chroma_4$ | $mfcc_6$ | $mfcc\Delta\Delta_6$ | $mfcc_{10}$ | $chroma_6$ |
| Features | $FFLpc_1$ | $lpcCoeff_1$ | $mfcc\Delta\Delta_4$ | $chroma_8$ | $chroma_3$ |
| | $chroma_5$ | $mfcc_1$ | $mfcc_{10}$ | $mfcc\Delta\Delta_3$ | $chroma_4$ |
| | $chroma_3$ | $lspFreq_4$ | $lspFreq_4$ | $chroma_5$ | $mfcc_7$ |
| | $mfcc_7$ | $mfcc\Delta\Delta_6$ | $lpcCoeff_1$ | $chroma_4$ | $chroma_5$ |
| | $mfcc_2$ | $mfcc_{10}$ | $chroma_{11}$ | $chroma_{11}$ | voiceProb |
| | $chroma_6$ | $chroma_{11}$ | $chroma_4$ | $mfcc_1$ | MCR |
| | $mfcc\Delta\Delta_6$ | $chroma_5$ | $lspFreq_2$ | $lspFreq_2$ | $chroma\Delta_0$ |
| | $chroma_8$ | $FFLpc_1$ | $mfcc_1$ | $lspFreq_4$ | $chroma_9$ |
| Class. C4.5 | 87.43% | 76.62% | 81.88% | 81.02% | 82.98% |
| Class. SVM | 88.6% | 76.46% | 83.22% | 81.84% | 82.77% |

consonants.

All DFs of the evaluated consonants share a set of 9 AFs, where only the ranking is different. That means that those 9 AF carry information not only for one DF, but for the whole class of sonorants, glides and laterals. This shared AFs can be useful when it comes to further applications, for example in phoneme recognition, when multiple DFs have to be predicted in parallel.

Interestingly, also several chroma non-standard features are among the highest ranked AFs qualifying them as candidates for the phoneme recognition feature pool. That is true for both their basic and their two complex derivations, the delta and delta delta regression.

With just 3 AFs the learning algorithms correctly classified about 88% for all DF. When taking into account the top 15 AFs it seems that they already describe the underlying data well enough to produce results close to the results on the complete AF set. Only the performance of the SVM with the polynomial kernel increases when all features are used, as it is less prone to over-adaption.

Privative features either exist for a phoneme or are undefined, which means they have no clear state when not active. Therefore, they might be activated also in segments for which they are not relevant. Thus expected they show classification results that are worse than for the binary features.

### 5.2. DFs of vowels

In the results of the binary DF evaluation, some AFs seem to have strong descriptive power. For example the first LSP, the third, fourth and sixth MFCC. Even though the top features are not as stable as within the examined consonant class and, based on the classification results, are overall not as descriptive as AFs of the evaluated consonants.

When comparing the examined vowels and consonants based on the AFs the consonant related DF values can, therefore, be predicted with higher accuracy.

As mentioned in the results, the classification results of the only privative feature on vowels is comparable to the results of binary features. We think that is due to the fact, that the descriptive power of the AFs for vowel DFs turned out to be generally smaller than for the examined consonants.

## 6. Conclusions

We have proposed a method to explore the connection of DFs, which are a high-level representation of articulatory characteristics of certain phonemes, to AFs, which are low-level features directly extracted from the signal. The results show that it is possible to select AFs that describe the DFs in a satisfactory way.

It turned out that at least for consonants a subset of AFs is of relevance for all examined DFs, which is a useful contribution for the feature pool design for phoneme recognition. The prediction accuracy of the DF of vowels is in general lower than for consonants and the highest ranked AFs are not as stable. The classification accuracy when regarding all 106 available AFs, increases only for the SVM, not for the decision tree. The decision tree seems to over-adapt the available training data, whereas the SVM can take advantage of the large number of features. Within the feature selection method, the inter correlation of the AFs is not taken into account. A principal component analysis could solve this problem and lead to a smaller set of features by keeping the descriptive power. In follow-up studies we will focus on dynamic AF patterns to model non-static segments as well as the acoustics of DF changes at segment transitions [3].

## 7. Acknowledgments

The work of the authors has been carried out within the CLARIN-D project [18] (BMBF-funded).

# 8. References


[1] R. Jakobson, C. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press, 1961.

[2] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, NY: Harper & Row, 1968.

[3] K. Stevens, "Toward a model of lexical access based on acoustic landmarks and distinctive features," *Journal of the Acoustical Society of America*, vol. 111, pp. 1872–1891, 2002.

[4] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, 2000.

[5] J. Koreman, B. Andreeva, and H. Strik, "Acoustic parameters versus phonetic features in ASR," in *Proc. ICPhS*, 1999, pp. 719–722.

[6] J. Hou, "On the use of frame and segment-based methods for the detection and classification of speech sounds and features," Ph.D. dissertation, Rutgers, The State University of New Jersey, 2009.

[7] E. Eide, "Distinctive features for use in an automatic speech recognition system," in *Proc. Eurospeech*, 2001, pp. 1613–1616.

[8] T. Fukuda, "Distinctive phonetic feature extraction for robust speech recognition," in *Proc. Acoustics, Speech, and Signal Processing*, 2003, pp. 25–28.

[9] K. J. Kohler, "Labelled data bank of spoken standard german the kiel corpus of read/spontaneous speech," in *ICSLP*, 1996.

[10] T. John, "Emu speech database system," February 2012. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:bvb:19-152839

[11] T. Hall, *Phonologie - Eine Einführung*, ser. De Gruyter Studienbuch. De Gruyter, 2000.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462. [Online]. Available: http://doi.acm.org/10.1145/1873951.1874246

[13] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE book," Online available (last date accessed 2013-05-29), May 2010.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[15] R. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993.

[16] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[17] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[18] "http://eu.clarin-d.de/index.php/en/," CLARIN-D web page.