Recognition of German Obstruents

Julia Hoelterhoff

Department of Linguistics
University of Konstanz, Germany
Julia.Hoelterhoff@uni-konstanz.de

10.21437/Interspeech.2005-770

Abstract

The aim of the present investigation was to find characteristics in the speech signal that allow distinguishing the German obstruents /pf/ and /ts/, /f/ and /s/ and /p/ and /t/. Two acoustic analysis techniques were chosen to separate the target phonemes according to place and manner of articulation: measuring duration distinguished them in manner (fricative, stop and affricate), whereas the calculation of the relative amplitude of particular frequency bands distinguished them according to place of articulation (labial and alveolar). The speech data was recorded in a natural dialog situation to gain features that are robust towards variation in the speech signal. The characteristics found will serve for the improvement of the "FUL" automatic speech recognition system [1]. The FUL speech recognition system is based on underlying phonological features providing robustness for all kinds of variations in the speech signal.

1. Introduction

The present paper focuses on the automatic distinction of certain acoustic cues of the affricates /pf/ and /ts/, the fricatives /f/ and /s/ and the stops /p/ and /t/, forming a complete labial-alveolar contrast in German over three different articulatory sound types (cf. [2]). Affricates are composed of a stop sequence and a fricative sequence (in the following referred to as closure and frication portion). Two linguistic theories are discussed controversially: whether affricates are mono-phonemic or bi-phonemic. Ladefoged [3] argues for the mono-phonemic view that "an affricate is simply the sequence of a stop followed by a homorganic fricative" (p. 53). The automatic computational distinction of these obstruents in the speech signal was reported as difficult by many authors investigating speech production. Several possibilities of acoustic analysis had been proposed, such as duration, formant frequency (e.g. [4], [5]) and energy analysis. Manner of articulation (affricates versus fricatives and stops) was described to be successfully discriminated by temporal measurements with different parameter values. Howell and Rosen [6] measured *rise time* for affricates and fricatives, by calculating the time interval between the onset of an affricate or fricative to its' amplitude maximum. They found that rise time was approximately twice as long for fricatives compared to affricates, independent of their position in the word (initial, medial or final). Therefore, the entire frication portion of a fricative is supposed to be longer than that of an affricate. Klatt [7] found the duration of a consonant cluster to be shorter compared to the same, isolated phonemes in combination with a vowel. From this follows that the stop and fricative portions are shorter when combined in an affricate compared to their duration occurring as a single fricative or stop, corresponding with the results of Repp, Liberman, Eccardt and Pesetzky [8].

For the distinction of place of articulation it was shown [9] that *relative amplitude* measured in certain frequency bands served to distinguish alveolar and labial phonemes in the respective groups of affricates, fricatives and stops. This metric was proposed in a wide-ranging study testing several methods of analysis to distinguish place of articulation in American English fricatives [10]. To measure *relative amplitude*, a DFT was taken at the vowel onset and the amplitude of the F3 (for the sibilants /s,z,ʃ,ʒ/) and F5 region (for the non-sibilants /f,v,θ,ð/) compared to the respective fricative region by subtracting both values.

Formant frequency analysis was not taken into account in this investigation because in [9] it was shown that relative amplitude in specific frequency bands was a more successful metric to distinguish place of articulation for the target obstruents in the present experiment. Nevertheless, it was shown in [9] that the logarithmic distance measure, as proposed by [11] for vowel distinction, also provided robust results.

To summarize the goals of this study, the first goal was to find robust acoustic cues in the speech signal that allow distinguishing them according to manner and place of articulation. The analysis techniques chosen to do so were temporal measurements to distinguish manner of articulation and relative amplitude to distinguish place of articulation, in a modified way as described in [10]. The secondary goal was to clarify the linguistic nature of affricates.

2. Experimental description

A data collection for the purposes of this experiment was conducted. Each speaker was recorded in different environments to achieve multiple kinds of background noise in order to gain only robust acoustic cues from the speech signal.

2.1. Participants

Four speakers of Standard Northern High German with no remarkable difference regarding their dialect were recorded. The two female and two male participants reported no impairment of speech. All subjects had an academic background and were not aware about the purposes of this investigation. The subjects were not paid.

2.2. Method and Materials

The temporal and amplitudinal acoustic properties of the *six* target obstruents /pf/ and /ts/, /f/ and /s/ and /p/ and /t/ were investigated in the environment of *seven* different vowels /a/, /e/, /i/, /o/, /u/, /oe/ and /ue/. Each target sound was recorded in *two* positions: word initially and medially (*6 x 7 x 2 = 84*).

For each of these constructions *three* respective example words were chosen (*84 x 3 = 252*). The test words were again recorded within *two* different tasks, a reading task, having the target word embedded into a sentence (always in the subject position of the nominal phrase and always preceded by the articles "die" and "der" (*the*), the latter spoken like /deɐ/ and therefore ending with a "schwa") and a sentence building task, in which the subjects had to build a sentence of two of the randomly mixed target words. For the latter task, the preceding sound or word was not predictable depending on the subjects' sentence. Altogether, 504 words were recorded per person (*252 x 2 = 504*). Since in Standard High German the voiceless fricative /s/ does not occur in word initial position, the voiced counterpart /z/ was used instead. It was supposed that the differences compared to the other phonemes should remain the same, apart from voicing.
The speech data was randomly presented to the subjects on the screen of a laptop (1. task: interval of 1 sec., 2. task: interval of 4 sec.). The data was recorded on a Sony DAT recorder using a Sony condenser microphone. The recording rate was 44.1 kHz being downsampled to 22 kHz on hard disc.

2.3. Analysis

The speech data was analyzed with respect to duration and amplitudinal acoustic properties in particular frequency bands. Segmentation of speech data was carried out with Kaylab Multispeech 2.5.1, LPC analysis with Matlab 6.5.1, statistic analysis with JMP 5.0.1. For the effect tests of the statistical analysis of variance (ANOVA) the following factors were chosen: *target* obstruent, vicinal *vowel*, *target x vowel* interaction, *task* were obstruent was recorded in, test *word* nested under factors *target* and *vowel*, *subject* set as random factor. The factors were the same for each analysis.

2.3.1. Temporal analysis

Temporal analysis was chosen to distinguish the target obstruents in manner of articulation. Two variables were taken into account: the duration of the whole phonemes and the duration of the phoneme segments – frication and closure portion. The nature of the segments allowed the discrimination of affricates and fricatives on the one hand (frication portion) and affricates versus stops on the other (closure portion). The segmental length of affricates was measured as such that affricates were broken down into the closure and frication portion whereby the frication portion started with the release of the closure (this should be kept in mind comparing the frication portion of fricatives and affricates, since the latter might contain parts of the stop). Since the aim of this investigation was to find robust acoustic cues for speech recognition, some cues that are generally known to influence duration were neglected to make the results of this investigation independent of them; such as overall speaking rate, syllabic and stress patterns. However, stress and syllable position was the same for most of the target sounds.

2.3.2. Energy analysis in frequency bands

The measurements referred to the relative amplitude (cf. [10]). Two Short-Time Fourier-Transformations were taken: one at the center of the vowel, the other at the center of the relevant obstruent portion, using a 30 ms Hamming window. For affricates and fricatives the window was placed around the center of the frication portion, for stops it was placed around the center of the release portion. The energy was taken from frequency bands splitted up into regions in steps of 1000 Hz, starting from 0 to 1000 Hz and ending in the band between 7000 and 8000 Hz (from now on 7-8 kHz). The relative amplitude was then calculated by taking the difference in dB at the center of the vowel and the center of the respective obstruent portion of the same frequency region. An ANOVA was calculated to distinguish the target sounds with respect to the place of articulation function within their group.

3. Results

The target obstruents were discriminated in two different ways, with respect to manner of articulation analyzing their duration and with respect to place of articulation analyzing relative amplitude in frequency bands.

3.1. Results duration

The main finding considering duration was that manner of articulation can be robustly distinguished by evaluating the phoneme length of the target obstruents in word initial and medial position.
Concerning the absolute length of the phonemes, the ANOVA revealed a main effect for the target obstruents [$F(5,871) = 162.86$, $p < .0001$] in word initial position as well as for those in word medial position [$F(5,869) = 563.74$, $p < .0001$]. *Post hoc* tests revealed significant differences in manner of articulation for all tested conditions (/pf/ and /ts/ versus /f/ and /s/ and versus /p/ and /t/). Word initially, the mean duration of affricates (above 200 ms, cf. dotted lines in Figure 1) is significantly longer compared to fricatives and stops being around 167 ms to 170 ms. The voiced initial fricative /s/ can be found at the bottom of Figure 1 with a mean duration of 98 ms averaged over all vowels.
The *post hoc* test for manner distinction in word medial position revealed for all tested conditions (same conditions as for word initial position) that affricates are a separate class compared to the respective fricatives and stops ($t = 0$). Affricates are nearly twice as long (cf. Figure 2, the two dotted lines represent the affricates) compared to fricatives and stops. Further, the mean duration of affricates is slightly longer in word medial position, whereas stops and fricatives behave vice versa, they tend to be longer in initial position (except initial voiced /z/).

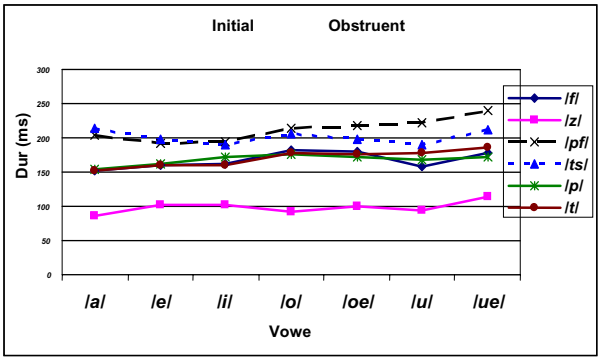

Figure 1: Comparison of mean duration of the target obstruents per vowel in word initial position.

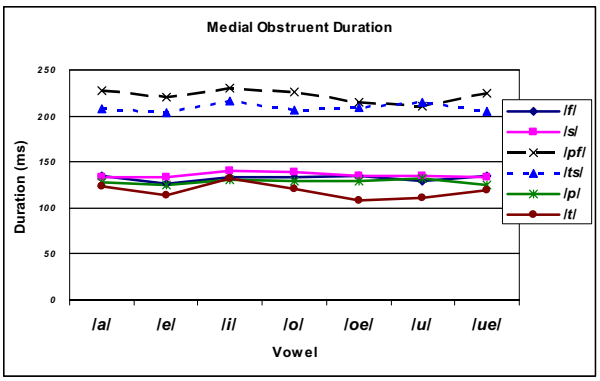

Figure 2: Comparison of mean duration of the target obstruents per vowel in word medial position.

The comparison of the closure portion of affricates and stops revealed a main effect for the variable *target* in word initial [$F(3,569) = 68.65$, $p < .0001$] and in medial position [$F(3,578) = 36.85$, $p < .0001$]. All *post hoc* tests on manner of articulation were highly significant, either in initial and medial position ($t < .0001$). Interestingly, the initial mean closure duration of stops (/p/ = 110 ms, /t/ = 106) was significantly longer compared to that of affricates (/pf/ = 60 ms, /ts/ = 82 ms) although the complete phonemes behaved vice versa. In word medial position, the mean duration of the closure portion was not following a regular pattern, /pf/ and /ts/ had nearly the same mean duration (84 ms and 83 ms) being shorter than /p/ (92 ms) and longer than /t/ (72 ms), although the medial affricates were nearly double as long compared to the respective stops. The evaluation of the frication portion of affricates and fricatives also revealed main effects in word initial [$F(3,570) = 152.28$, $p < .0001$] and in medial position [$F(3,577) = 8.39$, $p < .0001$] for the variable *target*. All *post hoc* tests were highly significant for word initial position ($t < .0001$). The frication portion of /f/ (corresponding to the complete fricative, mean duration 168 ms) was significantly longer compared to both affricates (/pf/ = 152 ms, /ts/ = 120 ms) whereas the voiced /z/ was by far shorter (98 ms). In word medial position the following results were achieved: /ts/ vs. /f/, ($t \leq 0.018$), /ts/ vs. /s/ ($t \leq 0.0004$), /pf/ vs. /f/, ($t \leq 0.0145$), /pf/ vs. /s/ ($t \leq 0.2045$). As for the closure portion in word medial position, no regular pattern distinguishing manner of articulation can be found. The mean of /pf/ (139 ms) and is longer than that of both fricatives /s/ (135 ms) and /f/ (132 ms) but also longer compared to /ts/ (126 ms). No gender differences were found.

3.2. Results relative amplitude in frequency bands

Measuring relative amplitude in frequency bands is a robust method to distinguish the investigated labial and alveolar target obstruents. Word initially and medially highly significant results were achieved demonstrating that relative amplitude in dedicated frequency bands distinguishes place of articulation. The ANOVAs of relative amplitude revealed main effects for the variable target in all frequency bands between 2-3 kHz to 7-8 kHz (all $p < .0001$, the frequency bands between 0-1 kHz and 1-2 kHz had less clear results):

Word initial position:

In Table 1, the results of the *post hoc* tests, evaluating the place of articulation contrast within the groups of fricatives, stops and affricates (labial versus alveolar), are displayed for all frequency bands that showed an effect for the variable *target*.

Table 1: *Post hoc* tests on place of articulation per frequency band, word initially

	2-3 kHz	3-4 kHz	4-5 kHz	5-6 kHz	6-7 kHz	7-8 kHz
/f/ vs. /z/	$t < .0001$	$t \leq .5734$	$t < .0001$	$t < .0001$	$t < .0001$	$t < .0001$
/p/ vs. /t/	$t < .0001$	$t < .0001$	$t < .0001$	$t < .0001$	$t < .0001$	$t < .0001$
/pf/ vs. /ts/	$t < .0001$	$t < .0001$	$t < .0001$	$t \leq .1859$	$t \leq .8983$	$t \leq .4405$

The findings correspond with [10] that in word initial position, the voiceless fricative /f/ has greater relative amplitude than the voiced /z/. In case of the present investigation this result was found for all tested frequency bands, as indicated in Table 1. The relative amplitude of labial /p/ (1.5 dB to 2.2 db) is smaller throughout all frequency bands compared to /t/ (8-10 dB). Word initial affricates were not discriminable in their higher frequency regions (between 5-6 kHz and 7-8 kHz). In all other frequency bands the relative amplitude of /pf/ is significantly smaller compared to that of /ts/ (in the other frequency bands, between 5-6 kHz and 7-8 kHz, the difference between /pf/ and /ts/ is minor 1 dB), following the tendency that labials have a smaller relative amplitude compared to the alveolar obstruents. Gender had no influence on these results.

Word medial position:

Table 2: *Post hoc* tests on place of articulation per frequency band, word medially

	2-3 kHz	3-4 kHz	4-5 kHz	5-6 kHz	6-7 kHz	7-8 kHz
/f/ vs. /s/	$t < .0001$	$t \leq .5734$	$t < .0001$	$t < .0001$	$t \leq .0001$	$t < .0003$
/p/ vs. /t/	$t < .0001$	$t < .0001$	$t < .0001$	$t < .0001$	$t < .0001$	$t < .0001$
/pf/ vs. /ts/	$t < .0001$	$t < .0001$	$t < .0001$	$t < .0001$	$t \leq .0024$	$t < .0040$

The results in Table 2 reveal that all of the tested place contrasts in word medial position are highly significant (apart from /f/ contrasted with /s/ measured in the frequency band between 3-4 kHz being probably due to a measurement artifact). In all of the place contrasts evaluated, the same tendency was demonstrated: labials have a significantly smaller relative amplitude in the frequency bands between 2-3 kHz to 7-8 kHz compared to their alveolar counterparts. Table 3 summarizes the results of the relative amplitude measured in word medial position in the respective frequency bands.

Table 3: Relative amplitude (dB) per obstruent per frequency band, word medially

	2-3 kHz	3-4 kHz	4-5 kHz	5-6 kHz	6-7 kHz	7-8 kHz
/f/	10.1	12.8	15.7	19.1	19.3	18.6
/s/	-5.9	-6.4	-6.3	-5.4	-5.6	-6.0
/p/	7.9	12.1	15.7	20.3	20.4	19.2
/t/	15.8	22.6	24.0	23.8	21.9	21.1
/pf/	9.3	9.2	7.8	7.2	5.1	4.3
/ts/	16.4	23.8	23.7	23.4	22.6	21.2

4. Conclusions

The present results indicate that temporal and amplitudinal acoustic analyses distinguish manner and place of articulation of the German obstruents /pf/ and /ts/, /f/ and /s/ and /p/ and /t/. Temporal measurements provide critical information to separate affricates from fricatives and stops. Affricates turned out to be significantly longer compared to fricatives and stops, in either word initial and medial position. The comparison of the phoneme segments also revealed significant differences: word initially, the closure and frication portions of stops and fricatives were significantly longer compared to the affricate equivalents – apart from initial /z/, being the only investigated phoneme characterized by voicing. Word medially, however, no regular pattern was found concerning the closure portion and frication portion analysis. Although the duration of the whole affricates is clearly longer compared to fricatives and stops (nearly twice as long in word medial position) the same tendency was not found for the segmental length. It seems that the segmental length is not linearly dependent on the duration of the whole phoneme as described in [8].

A systematic difference was found evaluating the duration of initial and medial whole affricates versus fricatives and stops. The medial affricates turned out to be nearly twice as long as the medial stops and fricatives whereas the same comparison in word initial position revealed a minor difference. A possible interpretation was that word initial affricates have a mono-phonemic nature whereas word medial affricates have a bi-phonemic nature. A possible conclusion was that historic derivation determines the role of German affricates (cf. [9]).

Relative amplitude in dedicated frequency bands was found to be a reliable function to distinguish place of articulation of the target obstruents. The frequency bands between 2-3 kHz up to 7-8 kHz proved to gain useful information for relative amplitude calculation. In either initial or medial word position, the labial obstruents turned out to have a smaller relative amplitude compared to the alveolar ones comparing them in their respective groups (affricates, stops, fricatives), apart from the initial fricative contrast between /f/ and /z/ were the voiced alveolar /z/ had the smaller relative amplitude.

5. Acknowledgements

This work was supported by the *Sonderforschungsbereich 471* funded by the *Deutsche Forschungsgesellschaft*. My thanks go especially to Henning Reetz for extensive discussion as well as to Aditi Lahiri for bringing in her valuable ideas. All errors are my own.

6. References

[1] Lahiri, A. and Reetz, H., "Underspecified recognition", in *Laboratory Phonology VII*, Gussenhoven, C. and Werner, N., Eds., Mouton, Berlin, 2002, pp. 637-675.

[2] Maddieson, I., *Patterns of Sounds*. Cambridge University Press, Cambridge, MA, 1984.

[3] Ladefoged, P., *A Course in Phonetics*. Harcourt Brace Jovanovich College Publisher, Fort Worth, 2001.

[4] Liberman, A. M., Delattre, P. C. and Cooper, F. S., "The role of selected stimulus variables in the perception of unvoiced stop consonants", *American Journal of Psychology, Vol. 65, 1952, pp. 497-516.*

[5] Stevens, K. N. and Blumstein, S. E., "The search for invariant acoustic correlates of phonetic features", in *Perspectives on the Study of Speech,* Eimas, D. and Miller, J. L., Eds., Lawrence Erlbaum Associates, Hillsdale, 1981, pp. 1-38.

[6] Howell, P. and Rosen, S., "Production and perception of rise time in the voiceless affricate/fricative distinction", *Journal of the Acoustical Society of America, Vol. 73, 1983, pp. 976-984.*

[7] Klatt, D. H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", *Journal of the Acoustical Society of America, Vol. 59, 1976, pp. 1208-1221.*

[8] Repp, B. H., Liberman, A. M., Eccardt, T. and Pesetsky, D., "Perceptual integration of acoustic cues for stop, fricative and affricate manner", *Journal of Experimental Psychology: Human Perception and Performance, Vol. 4, 1978, pp. 621-637.*

[9] Hoelterhoff, J. and Reetz, H., "Combing acoustic cues to distinguish German obstruents according to manner and place of articulation", (forthcoming).

[10] Jongman, A., Wayland, R. and Wong, S., "Acoustic characteristics of English fricatives", *Journal of the Acoustical Society of America, Vol. 108, 2000, pp. 1252-1263.*

[11] Miller, J. D., "Auditory-perceptual interpretation of the vowel", *Journal of the Acoustical Society of America, Vol. 85, 1989, pp. 2114-2134).*