Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German.

Cédric Gendrot[1] & Martine Adda-Decker[2]

1 LPP Université Paris Sorbonne Nouvelle CNRS UMR 7018 ILPGA
2 LIMSI-CNRS bât. 508, BP 133, 91403 Orsay cedex
cgendrot@univ-paris3.fr, madda@limsi.fr

Abstract

Formant values of oral vowels are automatically measured in a total of 50000 segments from four hours of journalistic broadcast speech in French and German. After automatic segmentation using the LIMSI speech alignment system, formant values are automatically extracted using PRAAT. Vowels with unlikely formant values are discarded (4%). The measured values are exposed with respect to segment duration and language identity. The gap between the measured mean Fi values and reference Fi values is inversely proportional to vowel duration: a tendency to reduction for vowels of short duration clearly emerges for both languages. These results are briefly discussed in terms of centralisation and coarticulation.

1. Introduction

The availability of large audio corpora and powerful automatic tools for alignment motivates the present study, which aims at observing the automatically determined formant values in oral vowels. French and German vowel systems will be compared and the obtained measures will be discussed as a function of segment length. By this experimental study we try to answer, at least partially, to several questions:
- Is it possible to automatically extract reliable formant values from an automatically aligned "spontaneous" speech corpus? To what extent? What proportion of vowels achieves formant values close to the awaited targets?
- How do formant values vary with respect to segment durations? Which vowels are most prone to variation?
- Which differences can be observed between the variations of the two languages?

This work aims at contributing to the establishment of values of formants and their variability, especially for French as they are hardly documented at the present time. Our interest goes towards the variations in terms of opening/closing (correlated to F1) and the frontness/backness (roughly correlated to F2) and we will discuss the possible reduction phenomena in terms of centralisation/coarticulation.

2. Corpora and methodology

The French and German corpora correspond to radio and TV journalistic shows: articulation, without being emphasized, remains quite distinct, so that speech can be understood by a broad audience. Such speech cannot be described as fully spontaneous, but rather as prepared speech: only few hesitations, repetitions, and word fragments are observed and syntactic structures often remain close to written language. Reduced vowel phenomena, which we are interested in throughout this study, are undoubtedly less present here than in more conversational-style spontaneous speech.

2.1. Corpora

The French corpus corresponds to approximately 2 hours of speech (15 men and 15 women) mainly extracted from broadcast news of France Inter, recorded and transcribed orthographically at the French CTA/DGA. The present study was carried out in the framework of the MIDL (Modelisations for Identification of Languages) project. MIDL partners were LIMSI-CNRS, LPP Paris3, CTA/DGA/GIP, Télécom Paris, EA1483 Paris3.
The German corpus corresponds to 2 hours of journalistic shows of ARTE (20 men and 10 women). These are resources which are available at LIMSI at research ends via various European projects (in particular LE-OLIVE and LE-ALERT). The acoustic quality of the German corpus is slightly poorer than that of the French corpus. For this reason we initially used more data in German (5 hours) than in French. In order to keep comparable proportions between both languages, the German corpus was limited to two hours (the comparison of average formant values measured on two and five hours did not show significant differences).

2.2. Automatic alignment

The LIMSI speech transcription system [1] was used for corpus alignment. Orthographical transcriptions being known a priori the alignment system is used to locate phone boundaries, to choose among potential pronunciation alternatives (in particular "liaisons" and schwa), and to discard silences, breath and other noise segments. Context-independent phone models are used for alignment. Whereas context-dependent (e.g. triphone) acoustic models produce better transcription performances (i.e. a lower word error rates), context-independent acoustic models are more reliable for phone boundary location
For technical reasons, the segmentation resolution is limited to 10ms and the minimum duration of a segment is 30ms. Labelling thus produced is not a phonetic, but rather a phonological or phonemic labelling (corresponding in most cases to standard word pronunciations). Formant measures then allow to evaluate the variations observed in the acoustic realisation of phonemes.

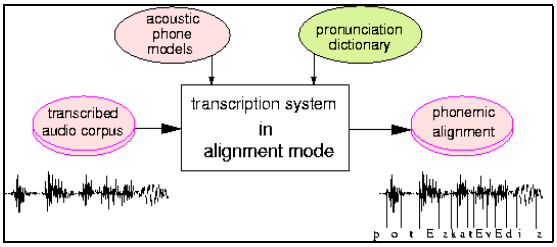


Figure 1: Overview of speech alignment : the audio stream is segmented and each segment is given an appropriate phonemic (or silence/breath/noise) label.

2.3. Automatic formant extraction

Formant extraction makes use of the Burg algorithm implemented in the PRAAT software [2]. The detection of amplitude peaks is determined in a band lower than 5kHz for male speakers and lower than 5,5kHz for females. Measurements were taken respectively at 1/3, ½, 2/3 of the vowel segment, and then averaged to provide a single value.

The interpretation of the extracted amplitude peaks as formants can raise controversy on a considerable number of segments: noises, too high fundamental frequency (voice of women and children), nasality... Two methodological safeguards are applied to prevent from errors:

(i) French nasal vowels were excluded from the study, German vowels being exclusively oral.

(ii) amplitude peak values are filtered in order to reject erratic items, with respect to the acoustics of the vocal tract. For each vowel, upper and lower formant value limits are given for the first three formants (cf. Table 4 in appendix): if one of the formants lies outside the specified ranges, the corresponding vowel segment is rejected.

Formant ranges were chosen in a broad way. A hundred visual checks for each vowel were carried out in order to reject as "errors" only severe formant detection problems and not the "deviating" values which might be due to contextual assimilation effects, to prosody or to speaker's characteristics for example. The formant values of a vowel depend not only on the speaker, but also on the articulation context (mainly left and right phonemes) and on the position of the vowel relatively to stressed and boundaries. The concept of "target undershoot", i.e. the non-realisation of the awaited target values is a well-known phenomenon, studied by many researchers [3,4]. However, to the best of our knowledge, large-scale corpus measures have not yet been carried out and compared across languages.

After this filtering, approximately 1000 vowels out of the 24000 oral French vowels were rejected (4% of segments rejected). The major part of these rejections corresponds to segments of very short duration (600 of the rejected segments have a duration smaller than 50ms). Listening to many of them shows that, at least for the shorter segments, the segmentation is not questionable. Other reasons may explain these rejections, in particular a partial or total devoicing of vowels, thus making formant detection more difficult (or even impossible) and potentially producing non-sense formant values. Similarly, when two formants of a vowel are close, especially in low frequency ranges, (which is the case for posterior closed vowels), the algorithm may detect only one formant instead of two, thus entailing a shift towards the higher order values. The /u/ is particularly prone to rejection, as all mentioned reasons may apply. For the /u/ this raises problems for the automatic control of formants (cf 3.1).

3. Analysis and results

Results are mainly given for French. The German corpus is used to check whether our observations hold for other languages than French.

3.1. Rejection rates of vowels

According to vowel identity: Table 1 indicates rejection rates obtained by the filtering described in 2.3. The overall rejection rate is about 4%. One can notice that rejection rates are highest for closed vowels (/i/, /y/, /u/). We may suggest that not only height would be involved in the rejection rate, but also the proximity of 2 formants for these vowels (namely F2 - F3 for /y/; F3 - F4 for /i/). /œ/[1] is also frequently rejected because of its duration (cf. below). /u/ is strongly rejected for reasons explained in section 2.3.

Table 1: Proportion of rejected segments as a function of segment identity (in %).

i	y	e	ɛ	a	œ	ø	ɔ	o	u
5	15	1	0.3	0.6	4	0.4	1	4.9	25

According to vowel duration: Table 2 indicates the rate of rejected segments for three duration intervals. Results show that filtering eliminates more frequently shorter segments. We hypothesise, in addition to the reasons invoked above for vowel rejection, that this can also be due to locally unclear pronunciations, where hypothesized segments are actually shorter than 30ms (cf. section 2.2.) and thus including in its segmentation a part of the preceding and/or following phonemes (consonants).

Table 2: Proportion of segments measured and rejected as a function of segment duration D (ms).

D duration	[30 - 50]	[60 - 80]	[90 – 110]
% proportion	39 %	38.5 %	22.5 %
% rejection	6.1 %	2.8 %	2.4 %

3.2. Variations in formants values according to duration

Average values of the vowels retained after filtering are shown in Table 5 (see appendix). As observed in Table 2 shorter segments are more prone to rejection than longer segments: this suggests that short vowels drive away from their F1 and F2 reference values as mentioned in the literature ([3]). To check this point statistically, ANOVAs with two factors were carried out, the two factors being "phoneme" and "duration". For vowel duration D, the specified intervals [D <= 50ms]; [60 <= D <= 80ms] and [90 <= D] allowed the use of these numerical values as nominals for statistics. Results show that segment duration has a significant effect on

[1] The schwa (/ə/) and /œ/ are merged, as they share the same representation in the speech recognition system with the same acoustic model for both.

the values of F1 [F(49.25) p<0.0001] and F2 [F(9.85) p<0.0001]. For German, observed variations as a function of vowel duration are shown in Figure 3 (see appendix): tendencies are globally similar to those observed for French. This may be interpreted for both languages as non-reached targets for the shorter vowels while the formants of longer vowels are close to references values as described in the literature (cf. table 3 for French, and [7] for German). More details and exact mean values of F1 and F2 formants according to duration intervals D can be found in [5].

Table 3: Formant reference values used for French as inspired from read isolated sentences [6].

V	i	y	e	ɛ	a	œ	ø	ɔ	o	u
$F1_{male}$	300	300	350	450	650	500	400	550	400	350
$F2_{male}$	2050	1800	1950	1700	1300	1450	1450	1050	900	850
$F1_{female}$	350	350	450	650	750	550	450	600	450	350
$F2_{female}$	2400	2050	2300	2000	1550	1650	1650	1200	950	850

3.3. Variations in formant values according to context

Duration of a segment is partly determined by the suprasegmental context. The "target undershoot" observed when looking at the figures gives a first impression of centralisation, as peripheral vowels shift to a more neutral position when shorter. However segmental context is an important parameter in determining vowels' spectral characteristics (as stated by Vaissière [9] for French).

Now if we consider that shortest vowels are more prone to coarticulation ([3]), i.e. taking spectral characteristics of the immediately preceding and following consonants, then short vowels would be more coarticulated and thus have a F2 transition pointing towards a locus corresponding to the place of articulation of neighbouring consonant (for example at approximately 600Hz for labial consonants - more specifically plosives - and 1800Hz for dentals [10]. For most vowels, this tendency is coherent with centralisation effects except for some specific cases such as / pap / : short [a] will not shift to a neutral position like other vowels as its second formant will lower compared to its target value (1444 Hz for male and 1677 Hz for female speakers).

In order to check this point for our data, we isolated in both corpora vowels with identical left-right segment considering place of articulation (for labials : p, b, m, f, v and for dentals : t, d, n, l). We can notice on figures 4 and 5 that short vowels in labial phonemic context are more closed but also more posterior (lowering of F2). This is interpreted as a higher coarticulation, rather than more centralisation, as more centralisation would lead F2 to move towards 1500 Hz instead of lowering. This interpretation is also valid for other contexts although formant movements cannot distinguish coarticulation from centralization. This tendency can be observed for French but not for German and this will be the next point of investigation of this study.

4. Discussion and conclusion

Although many points remain to be described and deserve a more in-depth investigation, the presented study allows to give first answers to the questions listed in the introduction. All the measured vowel formants of the corpus (minus 4% rejected by filtering) occupy the vocalic space in an organised manner : for segments which are longer than 90 ms, the automatically measured values describe a "triangle" which is very close to the reference values published for French in the literature. For shorter durations the vocalic triangle shrinks progressively resulting in "concentric" surfaces for both French and German (see Figures 2 and 3): the vocalic surfaces undergo a centripetal movement towards some central vowel position. The previous observations go in favour of the validity of automatic extraction from automatically aligned data, at least for the longer segments.

Comparing French and German, important differences in formant values could be measured for /y/ vowels between both languages. As German is a language with lexical accent, a more important tendency to reduction than for French could be expected (as suggested in [8]). However reductions are observed with a comparable degree in both languages, even if French high front vowels /i/, /y/, /e/ seem to be less prone to variation then German /i/, /y/, /Y/, /e/. This suggests that reduction is not an exclusively linguistic phenomenon, but admits also explanations of a physical or physiological nature.

5. Acknowledgment

The present study was funded by the French CNRS in the framework of an interdisciplinary research program (STIC-SHS *Société de l'Information* MIDL project).

6. References and Appendix

[1] Gauvain, J.L., Lamel, L. and Adda, G. (2002) The Limsi Broadcast News Transcription System, *Speech Communication,* 37(1-2):89-108.

[2] Boersma, P. and Weenink D. (1999) Praat, a system for doing phonetics by computer. Institute of Phonetic Sciences of the University of Amsterdam, report 132-182.

[3] Lindblom, B. (1963) Spectrogaphic study of vowel reduction, Journal of the Acoustical Society of America, Vol. 35, pp 1773-1781.

[4] Perrier, P. et Ostry, D.J. (1993) Dynamic modelling and control of speech articulators: application to vowel reduction. In E. Keller (Ed.), Fondamentals of speech synthesis and speech recognition, pp. 231-251.

[5] Gendrot, C. and Adda-Decker, M. (2004). *Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande*. Proceedings of the MIDL workshop, Paris.

[6] Calliope (1989) La parole et son traitement automatique, Collection Technique et Scientifique des Télécommunications, Ed. Masson.

[7] Rausch, A. (1972) Untersuchung zur Vokalartikulation im Deutschen. In Beitraege zur Phonetik von Heinrich Kelz und Arsen Rausch. IPK-Forschungsberichte (Bonn) 30, Hambourg, 35-82.

[8] Delattre, P. (1962) Comparing the prosodic features in English, German, Spanish, and French. Int. Rev. Applied. Linguistics I, 193-210.

[9] Vaissière, J. (1985) Etude des variations allophoniques de la voyelle /a/ et ses conséquences pour la reconnaissance automatique de la parole. In XIV Journées d'Etudes de la Parole, Paris.

[10] Delattre, P., Liberman, A., Cooper, F. (1955). *Acoustic loci and transitional cues for consonants*. JASA, 27(4).

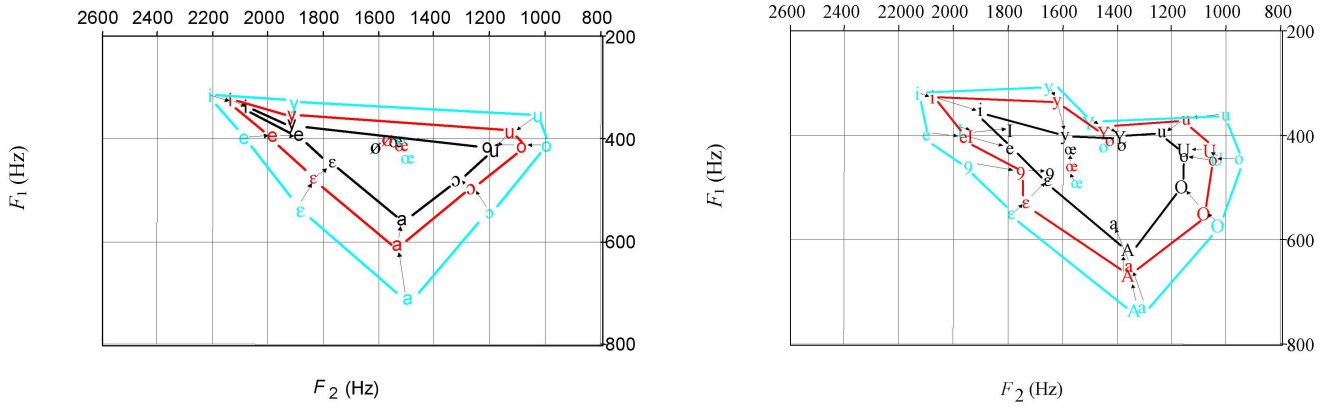
Table 4: Fi formant threshold values (minimum and maximum) in Hertz used for filtering.

French - male										
	i	y	e	ɛ	a	œ	ø	ɔ	o	u
F1	<750	<800	<800	<1000	<1000	<1000	<900	<900	<900	<900
F2	1500 - 2500	1300 – 2200	1100 - 2400	1200 - 2300	800 - 2300	800 - 2000	700 - 2000	600 – 1800	600 - 1600	400- 1500
F3	> 2000	> 1700	> 2000	> 2000	> 1800	> 2000	> 1700	> 1500	> 1500	> 1400
French - female										
F1	< 900	< 900	< 900	< 1100	< 1100	< 1100	< 1000	< 1000	< 1000	< 1000
F2	1600 – 3100	1400 -2800	1400 - 3000	1400 - 2700	900 - 2300	800 - 2400	700 - 2300	600 - 2000	600 - 1600	400-1500
F3	> 2500	> 1800	> 2200	> 2000	> 1900	> 2000	> 1800	> 2000	> 2100	> 1800

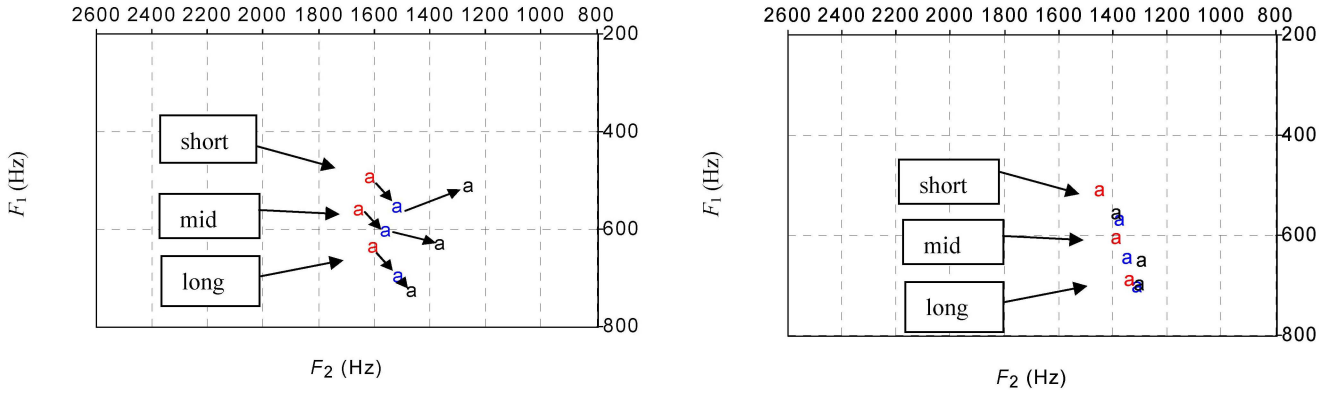
German - male															
	I	I	e	9	ɛ	a	A	o	O	u	U	y	Y	ø	œ
F1	< 750	< 900	< 800	< 1000	< 1000	< 1000	< 1000	< 900	< 900	< 900	< 900	< 800	< 900	< 900	< 1000
F2	1300-1500	1100-2500	1000-2500	1200-2300	1100-2300	800-2300	800-2300	600-1600	600-1800	400-1500	600-1800	1100-2200	700-2000	700-2000	800-2000
F3	> 2000	> 2000	> 2000	> 2000	> 2000	> 1800	> 1800	> 1500	> 1500	> 1400	> 1400	> 1700	> 1700	> 1700	> 2000
German - female															
F1	< 900	< 900	< 900	< 1100	< 1100	< 1100	< 1100	< 1000	< 1000	< 1000	< 1000	< 900	< 900	< 1000	< 1100
F2	1400-3100	1600-3100	1400-3000	1400-2700	1400-2700	900-2300	900-2300	600-1600	600-2000	400-1500	400-1600	1400-2800	700-2300	700-2300	800-2400
F3	> 2400	> 2400	> 2200	> 2000	> 2000	> 1900	> 1900	> 2100	> 2000	> 1800	> 1800	> 1800	> 1800	> 1800	> 2000

Table 5: Measured mean values of the first three formants of French according to vowel identity and speaker sex.

	i	y	e	ɛ	a	œ	ø	ɔ	o	u
$F1_{hommes}$	310	336	370	438	557	400	384	456	397	371
$F1_{femmes}$	348	371	423	526	685	436	420	528	438	404
$F2_{hommes}$	2005	1803	1850	1717	1444	1445	1474	1203	1041	1105
$F2_{femmes}$	2365	2063	2176	2016	1677	1643	1693	1347	1140	1153
$F3_{hommes}$	2784	2425	2545	2490	2438	2440	2405	2420	2477	2470
$F3_{femmes}$	3130	2745	2860	2800	2735	2715	2687	2743	2790	2742



Figures 2- 3: Measures mean average values of F1 and F2 for French (left) and German (right) vowels according to their duration. By ascending order (black [30 – 50ms], red [60 - 80], blue [90 - 110]). Male and female results are merged.



Figures 4- 5: Average values of F1 and F2 for French (left) and German (right) vowels according to duration (as fig. 2 & 3) and their left-right phonemic context (in order shown by arrows: red for dental, blue for all contexts, and black for labial).