INTER AND INTRA-SPEAKER VARIABILITY IN FRENCH: AN ANALYSIS OF ORAL VOWELS AND ITS IMPLICATION FOR AUTOMATIC SPEAKER VERIFICATION

Juliette Kahn[a,b], *Nicolas Audibert*[c], *Jean-François Bonastre*[a] *& Solange Rossato*[b]

[a]Laboratoire Informatique d'Avignon, Université d'Avignon et des Pays de Vaucluse, France;
[b]Laboratoire Informatique de Grenoble, Université de Grenoble, France;
[c]Laboratoire de Phonétique et Phonologie, Université Paris3-Sorbonne-Nouvelle/CNRS-UMR7018, Paris, France

juliette.kahn@univ-avignon.fr; nicolas.audibert@gmail.com; jean-françois.bonastre@univ-avignon.fr; solange.rossato@imag.fr

ABSTRACT

Intra and inter-speaker variability is studied as a way to better understand how voice can be used as biometric data. Formant values from 328,016 exemplars of the 10 French oral vowels uttered by 111 speakers were compared to estimate their speaker discrimination power. The vowels /œ/, /ɛ/ and /a/ appear to convey more idiosyncratic information than other oral vowels. A more comprehensive phonetic analysis is carried out for each speaker on 2 samples leading to either high or low discrimination performance when used in the Alize/spkDet SVS. However, no direct explanation can be drawn from phonetic measures to predict performance level.

Keywords: speaker verification, formant analysis, inter-speaker variability, intra-speaker variability

1. INTRODUCTION

Intra and inter-speaker variability in speech is an important issue in phonetics. For language description, its effects are often minimized by normalization [9]. On the contrary, inter-speaker variability is the basis of speakers' discrimination, which constitutes the focus of this article.

On the one hand, relationships between acoustic parameters and perceptual speaker discrimination have been studied over the last century [17]. According to [1], nasal vowels are more discriminant than oral vowels. A large inter-speaker variability in formant transitions was found by [16]. Analysis of F0 (means [18] and contours [7]), though not always discriminant, may improve speaker discrimination.

On the other hand, automatic speaker recognition systems have reached a rather high level of performance as shown by National Institute of Standards and Technology (NIST) evaluation campaigns [14]. However, system performance varies depending on different factors beginning with speech sample duration [8]. As automatic systems are mainly founded on stochastic methods, linking performance variations to acoustic descriptors of speech is not a straightforward task. Speech excerpts are parameterized by cepstral coefficients after automatic silence removal and parameter normalization [19]. Cepstral coefficient values provide information on phoneme spectral characteristics, while first order (delta) and second order (delta-delta) derivatives reflect short-term dynamic information. Longer-term information such as prosody and phoneme pronunciation order is not captured by cepstral coefficients. Although some systems [11] combine these statistical methods with sub-systems based on information such as prosody, nasality-related measurements or pauses, this information is not directly integrated in the modeling.

This paper focuses on the inter- and intra-speaker variability of French vowels, and its impact on automatic speaker verification systems. First, French vowel formants are analyzed in order to both try to predict the more relevant oral vowels for speaker discrimination and to measure the related intra-speaker variability. Second, the impact of intra-speaker variability on a speaker verification system is studied. A discussion concludes this paper.

2. EXPERIMENT 1

2.1. Formant analysis

Sentences read by French native speakers (64 female, 47 male) were selected from the BREF 120 corpus [12]. BREF sentences come from French newspapers and maximize phonetic coverage. A forced phonetic alignment was obtained using the open-source toolbox Speeral [13] and a manual adaptation of the phonetized lexicon to match actual realizations in the corpus. The resulting phonetic labeling associated each 10 ms frame with one of the 32 French phonemes.

In order to identify the oral vowels with the most idiosyncratic information, the first four formants were measured at the middle of the vowel for the 10 oral vowels of standard French /i, y, u, e, ø, o, ɛ, œ, ɔ, a/, adapting the LPC order according to the phoneme. All the measures were estimated with Praat [2]. 154,288 and 173,728 vowels were analyzed for male and female speakers respectively. Table 1 summarizes the number of occurrences of each vowel.

Table 1: Number of vowels analyzed from the BREF corpus, F1 to F4 values for the 10 oral vowels of standard French**,** and η^2 values for each vowel and speaker gender. Figures in formant values cells: bold=mean; normal=inter-speaker standard deviation; italics=intra-speaker standard deviation. Multivariate η^2 values indicate the magnitude of the speaker effect (ratio of explained variance) for MANOVA. All p-values are below 10^{-9}.

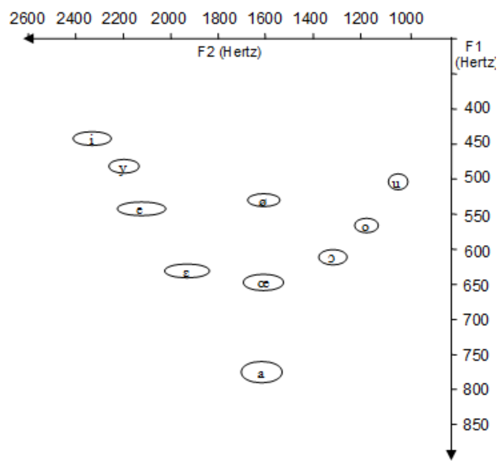
		/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
M	#	31,128	19,585	7,371	23,151	21,260	23,822	2,915	9,126	6,575	9,353
	F1	**602** 39-***94***	**498** 28-*90*	**509** 75-*115*	**434** 33-*114*	**487** 97-*142*	**384** 31-*113*	**509** 37-*60*	**521** 54-*93*	**434** 38-*113*	**412** 35-*126*
	F2	**1476** 49-*189*	**1759** 62-*138*	**1334** 225-*310*	**1951** 77-*125*	**1575** 97-*237*	**2363** 84-*207*	**1474** 58-*135*	**1312** 87-*239*	**1070** 40-*187*	**2223** 64-*187*
	F3	**2540** 113-*132*	**2595** 98-*132*	**2707** 144-*222*	**2694** 82-*144*	**2592** 117-*192*	**3040** 78-*187*	**2509** 102-*123*	**2572** 118-*172*	**2052** 89-*188*	**2890** 99-*223*
	F4	**3680** 163-*178*	**3686** 157-*189*	**3667** 124-*228*	**3709** 150-*183*	**3570** 121-*207*	**3662** 92-*191*	**3547** 144-*153*	**3580** 122-*180*	**2806** 80-*198*	**3542** 91-*181*
	Multivariate η^2	30%	30%	25%	29%	23%	16%	30%	26%	13%	15%
F	#	40,683	26,422	8,599	30,380	26,403	31,423	3,803	12,744	8,427	12,186
	F1	**708** 45-*98*	**571** 29-*82*	**505** 30-*87*	**481** 29-*79*	**471** 27-*72*	**383** 29-*90*	**587** 35-*54*	**551** 32-*78*	**443** 33-*100*	**422** 29-*113*
	F2	**1705** 85-*227*	**2021** 94-*181*	**1227** 49-*224*	**2229** 101-*173*	**1676** 67-*227*	**2409** 80-*130*	**1676** 84-*144*	**1382** 59-*222*	**1086** 41-*186*	**2260** 63-*193*
	F3	**2833** 147-*184*	**2911** 131-*162*	**2860** 141-*160*	**3006** 120-*151*	**2808** 126-*174*	**3021** 70-*169*	**2843** 148-*138*	**2846** 153-*155*	**1886** 68-*217*	**2876** 58-*176*
	F4	**3940** 212-*266*	**4004** 220-*265*	**3949** 144-*181*	**4050** 192-*242*	**3882** 135-*221*	**3620** 73-*191*	**3948** 172-*201*	**3939** 151-*195*	**2826** 72-*150*	**3597** 70-*183*
	Multivariate η^2	28%	29%	27%	26%	21%	15%	37%	28%	11%	10,00%

2.2. Inter- and intra-speaker variability

Mean values of formants 1 to 4 for each vowel, for male and female speakers, are also summarized in table 1 with inter- and intra-speaker standard deviation. Mean values match classical formant values for French oral vowels [5]. As shown by figure 1, inter-speaker variation depends on the vowel: while the inter-speaker standard deviation of F1 on /i/ is only 30 Hz, /ø/ has inter-speaker standard deviations of F1 close to 100 Hz.

Analyses of variance (ANOVA) with the speaker as fixed factor were run for each speaker gender and each oral vowel under consideration. The ANOVA allows us to compare the variability in formant values that can be attributed to inter-speaker vs. intra-speaker variation. Indeed, the distribution of Fisher's F function used in ANOVA represents the ratio of intra- and inter- variability for the factor of interest. In our analyses, a high F value therefore indicates a better discrimination among speakers. Since F and p-values cannot be directly compared, the discrimination power of different vowels is estimated by the size estimator η^2, which can be interpreted as the ratio of variance explained by the factor of interest [6].

Figure 1: Mean locations of the 173,728 exemplars of the 10 French oral vowels produced by female speakers. Ellipses represent inter-speaker standard deviation.



Vowels are firstly analyzed globally by using formants 1 to 4 as dependent variables in a multivariate ANOVA design (with Wilk's lambda criterion) with the speaker as fixed factor, for each vowel and each speaker gender. Though the speaker factor explains 21% to 37% of the variance for other oral vowels, which can be interpreted as a large effect [6], the effect of the speaker is weakest for the focal vowels /i/, /u/ and /y/. Overall, the mid vowels /œ/ and /ɛ/ and the low vowel /a/ appear to have the highest inter-speaker discrimination power. All Multivariate η^2 are presented in table 1.

Univariate variance analyses (performed on the same data subsets) using the speaker as fixed factor indicate that, for most vowels, effects are slightly stronger on F3 and F4 compared to F1 and F2. Focal vowels appear as the least variable among speakers for every formant, with an exception on F2 of /i/, especially for female speakers. However, this discrepancy should be interpreted with caution. Since F2 on /i/ is known to be weak, its automatic detection is more error-prone than for other oral vowels.

The intra-speaker variability is defined for each oral vowel and speaker gender as the mean of the standard deviations for each speaker. Intra-speaker standard deviation values are quite large. Their values fluctuate, for men, from 60Hz (/œ/) to 126Hz (/y/) on F1, and for F2 from 124 Hz (/e/) to 309 Hz (/o/).

3. EXPERIMENT 2

3.1. Methodology

3.1.1. Choice of the training excerpts

Using the Alize/SpkDet speaker verification system [10, 15] selected two samples of the same duration for each speaker of the BREF database: the one leading to the best performance (*Min*) and the one leading to the worst performance (*Max*) when the sample is used to build a speaker model. The performance is evaluated in term of Equal Error Rate [14], which represents the point where False Alarms ratio is equal to False Rejection ratio. Substantial intra-speaker variability was observed in system performance: from 0.9% of EER in *Min* set to 33% of EER in *Max* set.

3.1.2. Analyzed features

Based on the literature, a set of features known to be linked with idiosyncratic information were selected [7, 16]. For *Min* and *Max* sets, the features were extracted using the Praat software [2]. The studied feature set is composed of segmental information (phoneme numbers and trigrams of phonemes), formant values of the oral vowels and measures strongly influenced by co-articulation (area of the vocalic space [4], loci values [20]), F0, shimmer and jitter. Regarding the loci measures, the correlation *a* between the F2 value at 10% of the vowel and the F2 value at 50% of the vowel is measured as an estimation of co-articulation. These measures are computed for every oral vowel in bilabial, coronal or dorsal context. F0, shimmer and jitter values are analyzed for every oral vowel. The features values extracted from *Min* and *Max sets* are compared by paired t-tests for each vowel.

3.2. Results

3.2.1. Phoneme counts

Only the nasal consonants for female speakers ($p=0.025$) and the voiced fricatives ($p=0.037$) are significantly different in quantity between the two sets. No significant difference in the distribution of trigrams is found between *Min* and *Max* sets. Since samples are phonologically balanced, these results suggest that system performance variability might be better explained by differences in the intrinsic acoustic quality of speech segments.

3.2.2. Vowel quality

Only the /ɛ/ uttered by female speakers have F1 significantly different between *Min* and *Max* ($p<0.05$). Only the /e/ uttered by male speakers have F2 significantly different between both sets. No significant difference is found for F3. F4 are significantly different for /œ/ ($p<0.05$), /u/ ($p<0.05$) and /y/ ($p<0.05$) for female speakers. For male speakers, only F4 of /e/ are significantly different. No significant difference in vocalic triangle area is found ($p=0.7766$ for males, $p=0.9172$ for females). No significant difference in loci is found whatever the context ($p>0.07$).

3.2.3. F0 information

The mean F0 are not significantly different for female speakers ($p=0.790$) and slightly significantly different for male speakers ($p<0.01$). Similarly, the jitter values are not significantly different whatever the phonemes ($0.07612<p<0.9219$ for male speakers and $0.05083<p<0.9718$ for female speakers). For shimmer values, only the /i/ shows slightly significant differences ($p<0.05$) for female speakers when no significant difference is observed for male speakers ($0.1348<p<0.9796$).

It appears that the performance difference observed between *Min* and *Max* is not explained by suprasegmental and voice quality information.

4. DISCUSSION

In this paper, we tried to better understand the localization of speaker information in the speech signal. The analysis of inter- and intra-speaker variability constitutes a first step in this direction. Both inter- and intra-speaker variability are found to be large in the 328,016 French oral vowels analyzed. The comparison of speaker effect on formant values for oral vowels shows that, in French, vowels /a/, /ɛ/ and /œ/ convey more idiosynchratic information than the other oral vowels. Intra-speaker variability is an important factor for speaker verification systems. The formant values, the co-articulation information or the F0 were not able to explain this variability. However, [10] showed that the main difference between best and worst (training) speech excerpts is found on the cepstral values for all the phonemes except /v/. These differences, found on the cepstral level, are not explained by the analysis presented in this paper. Defining a confidence measure on automatic speaker verification results from the phonetic analysis of excerpts used as models therefore remains a challenge. The identification of relevant features for the modeling of idiosynchrasic information, and the evaluation of their impact on speaker verification is essential in voice biometric area, including forensic applications. Indeed, the ability to explain how the system makes the decision becomes crucial when important consequences are bound to this decision, as underlined in [3].

5. REFERENCES

[1] Amino, K., Sugawara, T., Arai, T. 2006. Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustic. Science. & Technology* 27.

[2] Boersma, P., Weenink, D. 2008. Praat: Doing phonetics by computer (Version 5.0.38) [Computer program]., *http://www.praat.org/*

[3] Bonastre, J-F., Bimbot, F., Boë, L-J., Campbell, J., Reynolds, D., Magrin-Chagnolleau, I. 2004. Authentification des personnes par leur voix: Un nécessaire devoir de précaution. *Journées d'Etudes de la Parole* 33-36.

[4] Bradlow, A., Torretta, G., Pisoni, D. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20(3-4), 255-272.

[5] Calliope. 1989. *La Parole et son Traitement Automatique*. Masson, Paris.

[6] Cohen, J.W. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

[7] van Dommelen, W.A. 1987. The contribution of speech rhythm and pitch to speaker recognition. *Language and Speech* 30(4), 325-338.

[8] Fauve, B., Evans N., Mason J. 2008. Improving the performance of text-independent short duration SVM and GMM-based speaker verification. *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

[9] Gray, A., Markel, J. 1976. Distance measures of speech processing. *Acoustic, Speech and Signal Processing* 24(5), 380-391.

[10] Kahn, J., Audibert, N., Rossato, S., Bonastre, J-F. 2010. Intra-speaker variability effects on speaker verification system. *Speaker and Language Recognition Workshop (IEEE Odyssey).*

[11] Kajarekar, S. 2008. Phone-based cepstral polynomial SVM system for speaker recognition. *Interspeech 2008 processing* 845-848.

[12] Lamel, L., Gauvain, J-L., Eskenazi, M. 1991. BREF, a Large vocabulary spoken corpus for French. *Eurospeech 91*

[13] Linares, G., Nocera, P., Massonie, D., Matrouf, D. 2007. The lia speech recognition system: from 10xrt to 1xrt. *Lecture Notes in Computer Science* 4629, 302-308.

[14] Martin, A., Greenberg, C. 2010. The NIST 2010 speaker recognition evaluation. *ICASSP processing* 1.

[15] Matrouf, D., Bonastre, J-F., Fredouille C., Larcher, A., Mezaache, S., McLaren, M., Huenupan, F. 2008. LIA GMM-SVM system description. *NIST SRE* Montreal, Canada.

[16] McDougall, K. 2006. Dynamic features of speech and characterization of speakers: towards a new approach using formant frequencies. *Speech Language and the Law* 13, 89-126.

[17] Mcghee, F. 1937. The reliability of the identification of voice. *Journal of .General .Psychology* 17, 249-27.

[18] Nolan, F. 2001. Speaker identification evidence: Its forms, limitations, and roles. *Proc. Law and Language, Prospect and Retrospect Conférence.*

[19] Reynolds, D.A. 2002. An overview of automatic speaker recognition technology. *Proc. IEEE Int Conf. Acoustic, Speech, and Signal Processing* Orlando, FL, 300-304.

[20] Sussman, H., Hoemeke, K., McCaffrey, H. 1992, Locus equation as an index of coarticultaion for place of articulation distinction in children. *Journal of Speech and Hearing Research* 35, 769-772.