On the Acoustic Characterisation of the Oral and Nasal Vowels of French

M. Eskenazi and J.S. Liénard
Orsay, France

1. Introduction

The manner in which the vowels of any given language have been characterised up to the present has not fundamentally changed from the ideas put forth by Peterson and Barney (1952). It has depended on the quality of phonologically and phonetically different vowels in that language as well as the relation of the vowels to one another in the acoustic dimension. Thus, the vowels of Japanese (there are five phonetically distinct entities (Komatsu et al., 1982)) could possibly be characterised by the type of information obtained from formant tracking. This implies that the phenomena of formants crossing one another, suddenly disappearing, or the apparition of a 'nasal' formant would be extremely unlikely to occur in Japanese. In French, there are generally considered to be ten oral vowels and three nasal vowels (for a typical Parisian speaker) – /a/, /ə/, /o/ /ɛ/, /e/, /i/ /y/, /u/, /œ/, /ø/, /ɛ̃/, /ɑ̃/, /ɔ̃/.

Despite this heavily populated vowel space, and due to the fact that there are no diphthongs or 'lax' vowels in French (so they may be considered to be 'steady state'), we shall show that these thirteen vowels may be sufficiently characterised by a relatively reduced amount of information in the frequency domain. This continues the work described in Liénard (1979). Articulatory representations, such as LPC, are not employed; rather an attempt is made to put to use the limited knowledge that presently exists on the manner in which the ear perceives sounds in the time and frequency domains.

After a description of the databases used, we explain the manner in which the speech signal was filtered and smoothed. A description of the simple statistics used to represent the dispersion of the vowels follows. The separation obtained, and the results of a first trial of unknown speaker vowel recognition are then presented as well as the explanation of a module that dynamically enlarges the base.

2. Databases

Several databases were made up: two to test the filtering, the spectral smoothings, and the statistical dispersions, and three to be used in the unknown speaker vowel recognition experiments.

A. Two test data bases

In order to explore different filtering and statistical approaches, and to confirm the dispersion results, two databases were constituted; one in context, and one of isolated vowels.

Both databases were recorded on a NAGRA IV S, with a BEYER M69N microphone. In the first there were 30 speakers, male and female, each recorded once (several repetitions of the list of sentences; the 'best' candidate for each vowel being retained). The frame sentence was, 'J'ai dit .... six fois' (/ʒɛdi .... sifwɑ/). The frame word always began with /t/, and ended either with one of the thirteen vowels mentioned above (example: thé /te/), or a consonant prolonging the duration of one of these vowels (example: thèse /tɛz/). The speech signal for each of the sentences was visualised, and the 50 ms portion to be used was indicated by hand (joystick).

The second, isolated vowel, database included ten speakers. These speakers pronounced the series of vowels three times at each recording session, and the 'best' candidate for each vowel was retained. Seven speakers (six male and one female) were recorded in two sessions (therefore two tokens of each vowel were present for each speaker); the other three speakers were recorded ten different times. The 50 ms portion of the vowel was obtained automatically : the starting point of the signal was detected, and the portion of interest was determined, after intelligibility tests, to begin at a fixed distance of 200 ms from this point.

B. Three recognition databases

For the unknown speaker recognition tests, three isolated vowel databases, alike in all ways except for the speakers included therein were constituted. Each included ten speakers, five male and five female, and were recorded on a REVOX B77, with the same microphone. There was only one recording session per person, but this time, of the five repetitions of the list that were requested, the three 'best' candidates for each vowel were retained. The 50 ms portions were excised in the same manner as above.

3. Filtering and Spectral Smoothing

Within 50 ms of speech and considering the possible use of a Hamming window, at least two, and as many as six, whole periods of the signal are present. This portion of vowel therefore seems to be correct as a base for obtaining a representative power spectrum. In all cases, the signal was preemphasized.

A. Filtering

In order to obtain the desired spectrum, two filtering methods were tested.

For the first database, 32 fourth order filters (characterised according to a Bark scale) were used. Satisfactory vowel groupings were obtained; however, to confirm these groupings, and to lighten the computing load, an FFT (translated to a Bark scale) was also tried. Results on the same database were quite comparable. The two filtering methods were then tried on the second database, and after results were found to again be quite comparable, the FFT was chosen for use in further tests.

B. Spectral Smoothing

Directly after the FFT and the transformation of the resulting spectrum into 32 values separated according to a Bark scale, the linear values were transformed to quasi-logarithmic (base two segment approximation) ones. The statistical treatment described below was used on this original unsmoothed 32-point spectrum to determine whether the smoothed spectrum conveyed different information in this context. Recognition tests of the first of the three recognition databases on known speakers showed comparatively high error rates.

Spectral smoothing was carried out with two goals in mind. First, the original spectrum still conveys a considerable amount of information, part of which may be considered to be redundant for our needs. Smoothing eliminates accidental peaks and valleys, but respects the general distribution of energy in the spectrum. Second, inherent differences in amplitude variations from one vowel to another are not taken into account. Smoothing may be carried out in several stages, allowing us to subtract an extremely smoothed spectrum from a less smooth one (both coming from the same original vowel). The result is a series of 37 values for each vowel (K = 1 to 37) representing the degree of curvature of the spectrum over a wide (~1000 Hz) range. The values are independent of amplitude and therefore an /a/ pronounced very softly will not be confused with an /i/, nor a loud /i/ with an /a/.

The combinations of smoothings found to give an optimal characterisation of the spectrum was:

1. three-point averaging, with the central value weighted at 2
2. 24-point weighted smoothing
3. 9-point straight averaging
4. substraction of 3. from 2.

Figure 1 illustrates the evolution of three spectra from their original forms to the results of 4, where only an indication of the degree of curvature (c) of the spectrum remains.

The original vowels were also treated in another manner: normalising amplitude before FFT, and then proceeding up to 2. Results of known speaker vowel recognition tests were slightly less satisfying than the procedure described above. Further tests are being carried out using this approach.

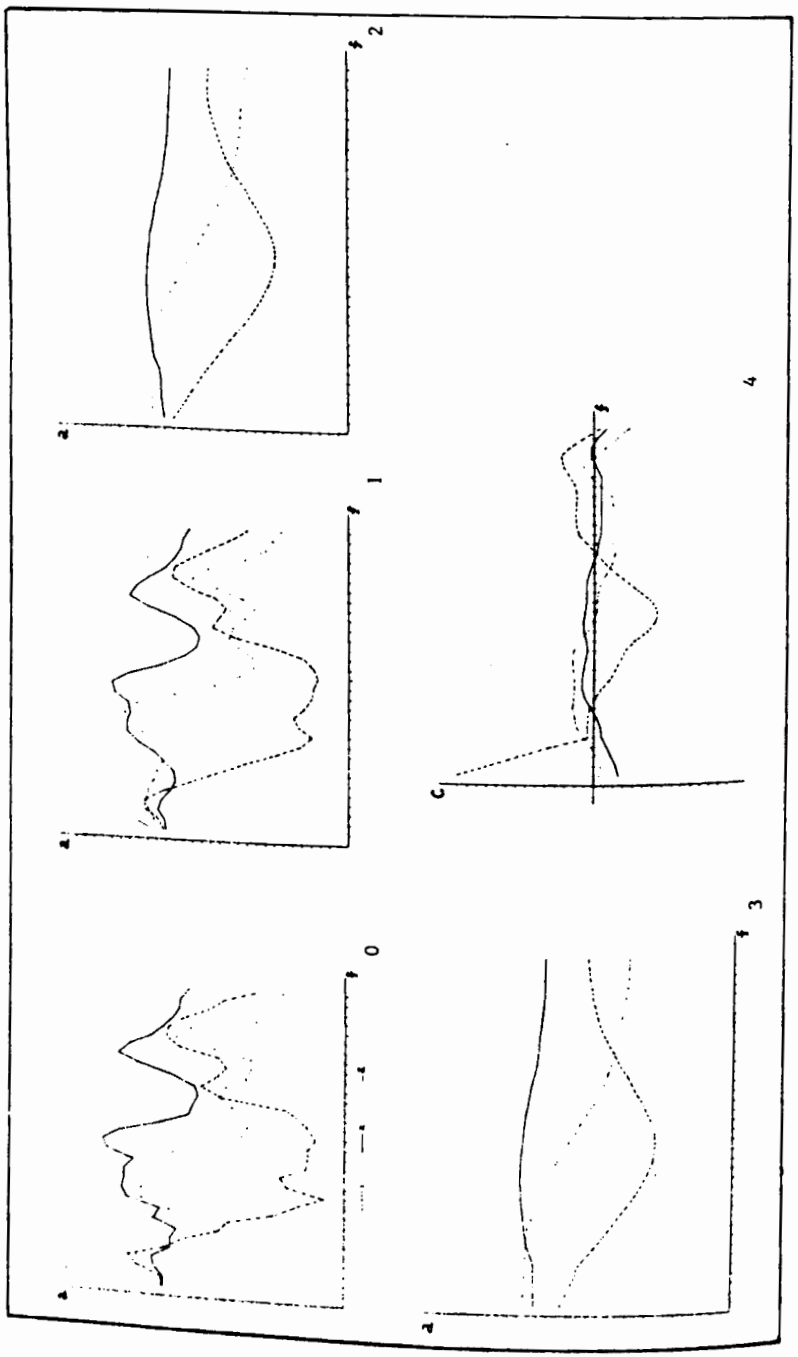

Fig. 1. Signal evolution from (0), the original power spectrum, to (4), amplitude-independent and smoothed.

4. Statistical Classification

The simple statistic tools chosen reflect the distance, at a given K, between the C values for different vowels (each 'vowel' now becoming a composite of all of the speakers in the given database).

First, the mean C value (m), and standard deviation (σ), for all the speakers for a given vowel were calculated at each K along the spectrum. We then calculated a dispersion value for all of the pairs of vowels at each K:

$$S(k, i, j) = \frac{|m_i - m_j|}{\sigma_i + \sigma_j}$$

A general indicator, (I(k)), of the dispersion of the vowels at each K, can be found by taking the mean value of the sum of the S(k,i,j) distances:

$$I(k) = 1/13^2 [\sum_i \sum_j S(k, i, j)]$$

Figure 2 shows the Ik values for the second and third databases: the higher the value, the better the general dispersion.

Figure 3 shows the general dispersion of the vowels at K = 30 for database 3. It may be noted that such vowel pairs as /a/ and /i/, and /a/ and /ɛ/ have

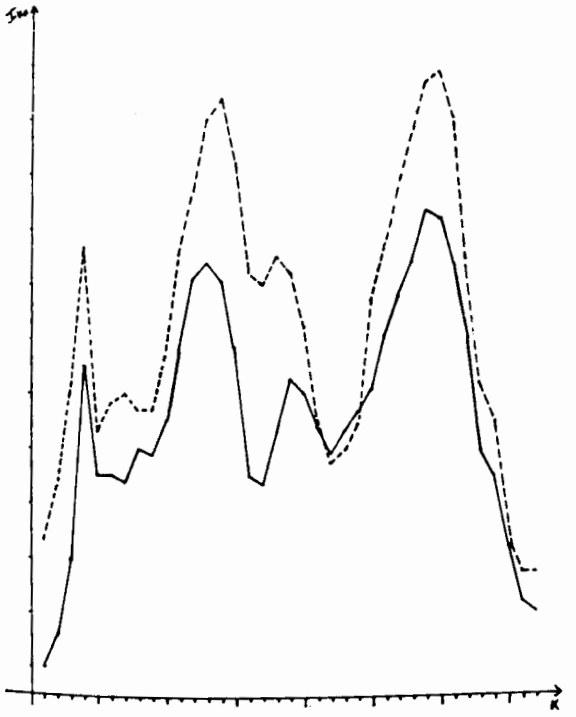

Fig. 2. Ik values for the second and third databases (dashes = database 2).

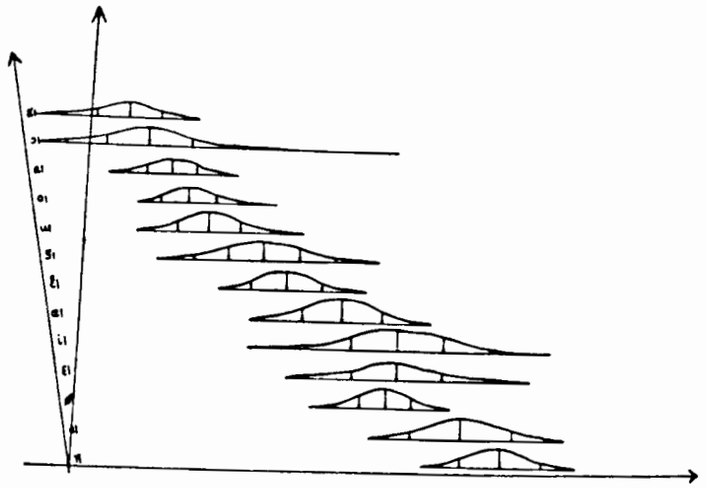

Fig. 3. General dispersion of the vowels at K = 30 (x axis = C, y Axis = vowels, z axis = N° of speakers).

well separated dispersions and cannot be confused with one another. At other values of K, the relations of the vowels among each other are different and, for example, at K = 14, /ɔ/ and /u/, and /u/ and /i/ are well separated.

The information contained in adjacent K's is quite similar, due to the smoothings. However, it would be interesting to see if an unknown vowel from an unknown speaker could be identified by superposition of certain Ks, or by trying to 'fit' the unknown spectrum against the 13 representative vowel curvature spectra obtained from the original database.

5. An Experiment in Vowel Recognition

Using the first of the three test databases to obtain the mean value and the standard deviation of each vowel at each K, we can filter and smooth a new unknown vowel from an unknown speaker (in the two other databases), and then compare its value at each K to the mean values of the 13 vowels in the base. The distance between the unknown, and a vowel mean value, at any given K is calculated as:

$$D_{uv} = \frac{|c_u - m_v|}{\sigma_v}$$

where u is the unknown vowel value at a given K, and v is the mean value of any one of the 13 base vowels. The shortest Duv distance should therefore indicate a likely vowel candidate. At K = 30, there is not enough information to make a decision on what the most likely vowel would be; however, a group of the three to four 'closest' candidates, almost always shows the vowel in question, and its closest 'neighbours' in the acoustic vowel space (for example, an unknown /a/ will often engender /o/, /ã/, /ɔ/, and /a/ as candidates).

We proceeded to test these distances on the whole of the spectrum. That is, taking the sum of the distances for all of the Ks for each of the 13 vowels, and seeing which vowel has the lowest cumulative distance from the unknown (the best 'total spectral fit'). Other tests were carried out to add up only half of the Ks (K = 2, K = 4, ... K = 36), one third, one fourth, one fifth, and one sixth of the Ks. Results are shown in Table I. Although so much of the information has been taken out of the signal, the correct vowel is closest to the unknown 45-50% of the time; it may be noted that when only half, one third, etc. of the Ks are used for comparison, the error rate does not change by more than a few points; one feature used for identifying oral and nasal vowels could only necessitate the retention of six m values and six σ values, New tests are presently being carried out on these same databases to see if the error rate decreases when only a few select Ks (the peaks shown in Figure 2, for example) are used for the cumulative distance calculations. Table II shows the confusion matrix for the total spectral fit.

Table I. Results of the vowel recognition experiment (as a percentage of 780 unknown vowels)

Test	Correct vowels proposed as place candidate (%)				
	1st	2nd	3rd	1st or 2nd	1st, 2nd, or 3rd
'Total Fit'	49.1	27.3	11.7	76.4	88.1
1/2 of the Ks	48.6	27.6	11.7	76.2	87.8
1/3 of the Ks	49.9	25.9	13.2	75.8	89.0
1/4 of the Ks	48.2	27.1	12.2	75.3	87.4
1/5 of the Ks	48.7	26.4	12.3	75.1	87.4
1/6 of the Ks	50.3	25.5	11.4	75.8	87.2

Table II. Confusion matrix for the total spectral fit

		confused with:												
		i	y	e	ɛ	ɛ̃	ø	œ	a	ã	ɔ	ɔ̃	o	u
unknown vowel	i	72	6											
	y	1	64	10	2		1							
	e	21	13	39	4							1		
	ɛ	1	4	17	45			1				10		
	ɛ̃				2	49		17	7		1	2		
	ø	1	3	1	6	1	46	13				7		
	œ		4		1	1	20	42	1	1	3	5		
	a					1			51	10	16			
	ã								5	32	38		3	
	ɔ								5	2	60	6	4	1
	ɔ̃						3	5			18	40	7	5
	o										24	4	46	4
	u									1	13	30	2	32

6. A Module to Dynamically Change the Statistics

The results of our recognition experiment may possibly be ameliorated in two ways; first, different manners of superposing the different Ks are being tried, second, the base of speakers is being enlarged to become more statistically representative, and therefore, used for unknown speakers with a decrease in error rates.

In order to enlarge the base for any vowel with any number of already known, or unknown speakers, we have developed a module that automatically changes the mean value, and the standard deviation value of the desired vowel. Information on the present status of the number of speakers in the base, the mean value, and the standard deviation, can be obtained at any time. Tests are now being carried out on the enlarged statistical base (now including all of the speakers in the second of the three test data bases).

7. Conclusions

We have shown that the ten oral and three nasal vowels of French may be characterized with quite a small amount of generalised information. This observation is in keeping with the work at LIMSI (Gauvain et al., 1983) which centers on the belief that the transitory parts of speech convey more information and therefore that less needs to be retained for the stable parts. Further work is being carried out on the enlarged statistical base (this base will soon be enlarged to encompass childrens' voices and whispered speech). Tests of the robustness of this representation are also being undertaken on vowel portions automatically taken from CV, VC, and C(C)VC syllables.

References

Cole, R., Alleva, F., Brill, S., Lasry, M., Phillips, M., Pilant, A., Specker, P., and Stern, R. (1982). FEATURE: Feature-based, speaker independent, isolated letter recognition. *Artificial Intelligence Conference*, Carnegie-Mellon University. Pittsburgh.

Gauvain, J.L., Liénard, J.S. and Mariani, J. (1983). On the use of time compression for word-based recognition. *IEEE-ICASSP*, Boston.

Komatsu, A., Ichikawa, A., Nakata, K., Asakawa, Y., Matsuzaka, H. (1982). Phoneme recognition in continuous speech. *IEEE-ICASSP*, Paris, pp. 883-86.

Liénard, J.S. (1979*a*). Sur quelques indices acoustiques des sons stables du français émis par plusieurs locuteurs. *9th International Congress of Phonetic Sciences*, Copenhagen.

Liénard, J.S. (1979*b*). Speech characterisation from a rough spectral analysis. *IEEE-ICASSP*, Washington, pp. 595-98.

Peterson, G., and Barney, H. (1952). Control Methods used in a study of the vowels. *J. Acoust. Soc. Am.*, **24**, 175-84.