The Amplitudes of the Peaks in the Spectrum as Acoustic Attributes of the Place of Articulation

Anna Esposito
Massachusetts Institute of Technology 02139 Cambridge (MA) USA
Universitá di Salerno, Dept. Fisica Teorica (SA) Italy

Abstract[1]
This work is devoted to the study of the properties of the sound spectrum at the release of Italian stop consonants in vocalic contexts. The aim is to check if the amplitudes of the peaks in the spectrum can be used as acoustic attributes of the place of articulation of the consonants. Moreover, different measurements have been performed in order to define what of measure retains more information about peak amplitudes.

Materials and procedures

The recording and measurements were made in the Research Laboratory of Electronics, Speech Communication Group, MIT, Cambridge, Usa. The materials consists in VCVC utterances produced by seven adult Italian speakers (three females and four males) in a sound-treated room and recorded on a high-quality magnetic tape recording system. The speakers were selected from different parts of Italy. The utterances are embedded in a carrier phrase: *Prendi VCVC se vuoi (Take VCVC if (you) want)*. The measurements were made for the consonant between vowels. Data have been collected for all the official Italian vowels embedded in stop contexts. However, the results reported in this paper come from the analysis of the stop consonants in the context of the vowel /*i*/. A more detailed description of the speech materials and the procedures can be found in Esposito and Stevens [2]. The spectral representations used include a DFT spectrum, a smoothed version of the DFT, a spectral averaging method. The analysis window (Hamming window) was set to 3.1 msec for each measurement. The spectrum at the release of each consonant, the averaged spectrum during the first 4 msec (for /*b, d, g*/) and 10 msec (for /*p, t, k*/) after the release and the *k*-averaged[2] spectrum were computed. All spectra are preemphasized, i.e. spectra of the first difference of the waveform are calculated. Moreover, the spectral amplitudes were also enhanced by changing an overall spectral gain control parameter. The amplitudes of the maximum peaks in the frequency ranges of 1-3kHz, 3-5kHz, 4-6kHz, 5-7kHz, 0-2kHz, and 0.8-1.5kHz were measured from cursor amplitude readouts via mouse position placed on the spectrum display. The spectrum display shows, superimposed, both the smoothed spectra and the DFT spectrum. However the peak amplitudes were measured only on the DFT spectrum.

The amplitude attributes

The peaks amplitudes measured in the different frequency ranges described above were compared in order to identify properties that can be useful to discriminate the place of articulation of each consonant. Initially averages of the maximum peak amplitudes in different frequency ranges were computed. However, even though some of these averages differ significantly from one consonant to another, the standard deviations were high and they overlapped. This effect is mostly due to the variability of the peak amplitudes among the speakers. For this reason we decided to exclude these measures and we start to look to the amplitudes of the maximum peaks in specified frequency ranges compared to the amplitudes of the maximum peaks in other frequency ranges. This comparison seemed more reasonable to us because it is possible to reduce the amplitude variability among speakers and repetitions. We carried out several attempts, comparing the maximum peak amplitudes in some frequency ranges with the maximum peak amplitudes in some other frequency ranges or comparing the differences between the maximum peak amplitudes in the different frequency ranges examined. In each attempt we defined a set of acoustic features based on these comparisons and tested this set of features on the consonants in order to verify if it accepted the consonant under examination and rejected the others. The final results of this trial and error process are the following set of acoustic attributes for each place of articulation:

Velar amplitude attributes:
a1) The differences between the maximum peak in the 0-2kHz and the maximum peak in the 4-6kHz frequency ranges must be lower than 2dB;
b1) The differences between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must be greater or equal to 9dB;
c1) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 4-6kHz frequency ranges must be greater or equal to 0dB;
d1) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 5-7kHz frequency ranges must be greater or equal to 0dB.

Labial amplitude attributes:
a2) The differences between the maximum peak in the 0-2kHz and the max-

[1]Supported by CNR-IIASS contratto quinquennale and INFM Salerno University. Acknowledgements goes to Maria Marinaro, Carmen D'Apolito and Kenneth N. Stevens for their comments and suggestions.

[2]This spectrum was computed measuring, for each voiceless consonant, the VOT length. Then the cursor was placed on the waveform at the temporal sampling point corresponding to the half of the VOT length and the spectrum averaged on 5 msec to the left and 5 msec to the right of this sampling point was computed. We call this spectrum the *k*-averaged spectrum because *k* is the command to compute it. *k* is a parameter of the analysis tool (Klatt tools [3]) previously set to 150 samples, corresponding to 10 msec of signal duration, at the sampling rate of 16.000 Hz.

Table 1: *Amplitude feature-matching results for velar consonants. The entries give the mean percentage of utterances of each consonant (based on 21 utterances of each consonant, occurring in /i/ vowel environment, and obtained from seven speakers) that were correctly accepted or rejected by the set of acoustic features defined above.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/k/ 57.1	/p/100 /t/95.2
/g/ 90.4	/b/100 /d/85.7
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/k/ 95.2	/p/100 /t/95.2
/g/ 80.9	/b/90.4 /d/76.2
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/k/ 95.2	/p/95.2 /t/95.2

Table 2: *Template-matching results obtained using the Blumstein and Stevens compact template.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/k/ 33.3	/p/ 95.2 /t/ 90.4
/g/ 66.6	/b/ 57.1 /d/90.4
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/k/ 47.6	/p/ 61.9 /t/ 85.7
/g/ 52.3	/b/ 47.6 /d/66.6
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/k/ 65	/p/ 61.9 /t/ 76.2

Table 3: *Amplitude feature-matching results for labial consonants.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/p/ 33.3	/k/ 95.2 /t/ 100
/b/ 66.6	/g/ 100 /d/95.2
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/p/ 90.4	/k/ 100 /t/ 90.4
/b/ 57.1	/g/ 100 /d/100
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/p/ 90.4	/k/ 100 /t/ 100

Table 4: *Template-matching results obtained using the Blumstein and Stevens diffuse-falling template.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/p/ 38.1	/k/ 95.2 /t/ 66.6
/b/ 57.1	/g/ 90.4 /d/85.7
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/p/ 61.9	/k/ 100 /t/ 90.4
/b/ 66.6	/g/ 90.4 /d/80.9
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/p/ 52.3	/k/ 100 /t/ 85.7

imum peak in the 4-6kHz frequency ranges must be greater than 1dB;

b2) The differences between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must be lower than 9dB;

c2) The differences between the maximum peak in the 1-3kHz and the maximum peak in the 5-7kHz frequency ranges must be greater than 8dB.

Alveolar amplitude attributes:

a3) The differences between the maximum peak in the 1-3kHz and the maximum peak in the 5-7kHz frequency ranges must be lower than 9dB;

b3) The differences between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must be lower than 10dB;

c3) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 4-6kHz frequency ranges must be lower than 0dB;

d3) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 5-7kHz frequency ranges must be lower than 9dB.

Table 5: *Amplitude feature-matching results for alveolar consonants.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/t/ 90.4	/k/ 80.9 /p/ 33.3
/d/ 80.9	/g/ 100 /b/90.4
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/t/ 85.7	/k/ 95.2 /p/ 90.4
/d/ 66.6	/g/ 90.4 /b/90.4
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/t/ 95.2	/k/ 90.4 /p/ 95.2

The set of acoustic attributes defined above are the same both for voiced and voiceless consonants. However, for voiced consonants we have to changes the 0-2kHz and 4-6kHz frequency ranges to 0.8-1.5kHz and 3-5kHz respectively. These frequency changes can be justified considering that in order to allow vocal-fold vibrations during the production of voiced consonants the larynx is lowered, the pharynx is expanded and the walls of the vocal tract are compressed. This could cause small shifts in the vocal tract resonances such as a lowering in frequency.

Results

We used the set of features defined above and the templates defined by Blumstein and Stevens [1] and we tested their discrimination performances. We obtained the results reported in the tables. These preliminary

results show that the amplitudes of the peaks in the spectrum computed during the first 10 msec after the release and in the *k*-averaged spectrum can be used to discriminate among the voiceless consonants /p, t, k/ (see tables 1, 3, 5). What is mostly useful to discriminate /p/ from /k/ is the property *a2* (even though also *b2* plays an important role) whereas *c2* is mostly useful to discriminate /p/ from /t/. The properties that allow to discriminate /k/ from /p/ are *a1, b1, d1*, whereas /k/ is successfully distinguished from /t/ by *b1*. The opposite of *a2 (a3)* is mostly used to discriminate between /t/ and /p/ and the opposite of *b1 (b3)* is used to discriminate between /t/ and /k/. This information can be used to define an automatic algorithm which discriminates successfully among /p, t, k/. Using the Blumstein and Stevens templates on the same data (see tables 2, 4, 6) the discrimination performances are less good in most of the cases. This result is expected in the case of the alveolars because of the different point of constriction of Italian /t, d/ with respect to American /t, d/. However, the results for labials and velars does not seem to be better suggesting some language specific influences on the gross shape of the spectrum.

In the case of voiced consonants, the set of attributes defined above can be used to identify /g/ and to discriminate /g/ from /b, d/ (at the release). However, for /b, d/ similar information does not identify the two consonants, even though they discriminates /b/ from /g, d/ and /d/ from /b, g/. In such cases, information about formant transitions is required. The voicing, which is always present in Italian, causes pressure fluctuations that lead to variability in the peak amplitudes.

With regard to the particular spectra computed it is possible to say that, in the case of voiceless consonants, the better performances of the acoustic attributes defined above and the Blumstein and Stevens template are obtained when the spectra during the first 10 msec after the release and the *k* averaged spectra are used for the comparisons. These spectra seem more useful to retain information about amplitude features. The spectra at the release retain more information about the amplitude attributes of voiced consonants.

Table 6: *Template-matching results obtained using the Blumstein and Stevens diffuse-rising template.*

Spectrum at release	
Correct Acceptance	Correct Rejection
/t/ 42.8	/k/ 47.6 /p/ 38
/d/ 52.3	/g/ 61.9 /b/80.9
Averaged Spectrum	
Correct Acceptance	Correct Rejection
/t/ 71.4	/k/ 71.4 /p/ 85.7
/d/ 42.8	/g/ 57.1 /b/85.7
***k*-Averaged Spectrum**	
Correct Acceptance	Correct Rejection
/t/ 61.9	/k/ 70 /p/ 90.4

These results are restricted to the consonants in the /i/ vowel environment. We will test the set of acoustic attributes defined in this paper to the consonants in the other vowel environments. We expect that there will be changes in their definition in order to improve their performances in the other vowel environments.

References

[1] S.E. Blumstein, K.N. Stevens, 1979, *Acoustic Invariance in Speech Production:...*, JASA, Vol. 64(4), 1001-1017.

[2] A. Esposito, K.N. Stevens, 1994, *Note on Italian Vowels:...*, (in press on MIT Speech Com. Work. Prog.).

[3] D.H Klatt, 1984, *MIT SpeechVax User's Guide*, Copyright 1984 by Dennis H. Klatt.