FORMANT BANDWIDTH AFFECTS THE IDENTIFICATION OF COMPETING VOWELS

Alain de Cheveigné
CNRS - IRCAM, France, and ATR-HIP, Japan

ABSTRACT
Formant bandwidth is known to have little effect on vowel quality. This paper shows that it has a strong effect on mutual masking between vowels. Subjects presented with stimuli consisting of pairs of synthetic vowels were requested to report one or two vowels for each stimulus. Identification rates were calculated independently for both vowels in the stimulus. Vowels had either the same or different fundamental frequencies. Their RMS amplitudes differed by 5, 15 or 25 dB. Formant bandwidth of each vowel was either twice or half its standard value. Identification of a target vowel was best when: (1) its RMS amplitude exceeded that of its competitor, (2) its formants were narrow, (3) formants of the competitor were wide, and (4) F0s were different. These effects were approximately orthogonal. A narrow-bandwidth voice is thus more resistant to masking, and a stronger masker, than a wide-formant vowel.

1. INTRODUCTION
Formant bandwidth is known to have little effect on the quality or intelligibility of isolated vowels [7,8]. However, if two vowels are in competition, as when two people speak at the same time, one can imagine that formant bandwidth might affect identification in several ways. For a given RMS amplitude, a formant attains locally a higher spectrum level if it is narrow than wide (Fig. 1), so a narrow-formant vowel might be more resistant to noise. On the other hand, interformant valleys are deeper when formants are narrow than wide, which might allow a competitor's formant peaks to emerge more easily. A narrow-formant vowel might thus be a less severe masker. There is therefore ample reason to suspect that formant bandwidth might affect identification of vowels in competition.

2. METHODS
The general methods are described in detail in [2,3]. In brief, stimuli were "double vowels" obtained by adding waveforms of two single vowels with amplitude ratios ranging from 5 to 25 dB in 10 dB steps. Single vowels were 5-formant synthetic Japanese vowels (/a/, /e/, /i/, /o/, /u/), with formant bandwidths one half or twice "normal", and fundamental frequencies (F0) of either 124 or 132 Hz, allowing F0 differences (ΔF0) of 0 and 6%. Formant frequencies were taken from [6], and "normal bandwidths" from [1]. Single vowels were synthesized with 270 ms durations, including 20 ms raised-cosine onsets and offsets. They had a "random" starting phase spectrum, the same for all vowels.

Single-vowels waveforms were scaled to a standard RMS amplitude after synthesis. To obtain a double vowel, two vowels were paired, one was scaled by a factor (5, 15 or 25 dB), both were added, and the sum was scaled to a standard RMS amplitude. The stimulus set included (20 pairs) x (3 amplitude ratios) x (2 ΔF0s) x (2 F0 orders) x (4 bandwidth combinations) = 960 stimuli. Stimuli were presented diotically via earphones. Sound pressure level varied between 63 and 70 dB(A) according to the stimulus

1. INTRODUCTION
Formant bandwidth is known to have little effect on the quality or intelligibility of isolated vowels [7,8]. However, if two vowels are in competition, as when two people speak at the same time, one can imagine that formant bandwidth might affect identification in several ways. For a given RMS amplitude, a formant attains locally a higher spectrum level if it is narrow than wide (Fig. 1), so a narrow-formant vowel might be more resistant to noise. On the other hand, interformant valleys are deeper when formants are narrow than wide, which might allow a competitor's formant peaks to emerge more easily. A narrow-formant vowel might thus be a less severe masker. There is therefore ample reason to suspect that formant bandwidth might affect identification of vowels in competition.

2. METHODS
The general methods are described in detail in [2,3]. In brief, stimuli were "double vowels" obtained by adding waveforms of two single vowels with amplitude ratios ranging from 5 to 25 dB in 10 dB steps. Single vowels were 5-formant synthetic Japanese vowels (/a/, /e/, /i/, /o/, /u/), with formant bandwidths one half or twice "normal", and fundamental frequencies (F0) of either 124 or 132 Hz, allowing F0 differences (ΔF0) of 0 and 6%. Formant frequencies were taken from [6], and "normal bandwidths" from [1]. Single vowels were synthesized with 270 ms durations, including 20 ms raised-cosine onsets and offsets. They had a "random" starting phase spectrum, the same for all vowels.

Single-vowels waveforms were scaled to a standard RMS amplitude after synthesis. To obtain a double vowel, two vowels were paired, one was scaled by a factor (5, 15 or 25 dB), both were added, and the sum was scaled to a standard RMS amplitude. The stimulus set included (20 pairs) x (3 amplitude ratios) x (2 ΔF0s) x (2 F0 orders) x (4 bandwidth combinations) = 960 stimuli. Stimuli were presented diotically via earphones. Sound pressure level varied between 63 and 70 dB(A) according to the stimulus

1. INTRODUCTION
Formant bandwidth is known to have little effect on the quality or intelligibility of isolated vowels [7,8]. However, if two vowels are in competition, as when two people speak at the same time, one can imagine that formant bandwidth might affect identification in several ways. For a given RMS amplitude, a formant attains locally a higher spectrum level if it is narrow than wide (Fig. 1), so a narrow-formant vowel might be more resistant to noise. On the other hand, interformant valleys are deeper when formants are narrow than wide, which might allow a competitor's formant peaks to emerge more easily. A narrow-formant vowel might thus be a less severe masker. There is therefore ample reason to suspect that formant bandwidth might affect identification of vowels in competition.

2. METHODS
The general methods are described in detail in [2,3]. In brief, stimuli were "double vowels" obtained by adding waveforms of two single vowels with amplitude ratios ranging from 5 to 25 dB in 10 dB steps. Single vowels were 5-formant synthetic Japanese vowels (/a/, /e/, /i/, /o/, /u/), with formant bandwidths one half or twice "normal", and fundamental frequencies (F0) of either 124 or 132 Hz, allowing F0 differences (ΔF0) of 0 and 6%. Formant frequencies were taken from [6], and "normal bandwidths" from [1]. Single vowels were synthesized with 270 ms durations, including 20 ms raised-cosine onsets and offsets. They had a "random" starting phase spectrum, the same for all vowels.

Single-vowels waveforms were scaled to a standard RMS amplitude after synthesis. To obtain a double vowel, two vowels were paired, one was scaled by a factor (5, 15 or 25 dB), both were added, and the sum was scaled to a standard RMS amplitude. The stimulus set included (20 pairs) x (3 amplitude ratios) x (2 ΔF0s) x (2 F0 orders) x (4 bandwidth combinations) = 960 stimuli. Stimuli were presented diotically via earphones. Sound pressure level varied between 63 and 70 dB(A) according to the stimulus

1. INTRODUCTION
Formant bandwidth is known to have little effect on the quality or intelligibility of isolated vowels [7,8]. However, if two vowels are in competition, as when two people speak at the same time, one can imagine that formant bandwidth might affect identification in several ways. For a given RMS amplitude, a formant attains locally a higher spectrum level if it is narrow than wide (Fig. 1), so a narrow-formant vowel might be more resistant to noise. On the other hand, interformant valleys are deeper when formants are narrow than wide, which might allow a competitor's formant peaks to emerge more easily. A narrow-formant vowel might thus be a less severe masker. There is therefore ample reason to suspect that formant bandwidth might affect identification of vowels in competition.
factors ($\Delta F_0 = 0, 6\%$) x (target bandwidth = half, twice) x (competitor bandwidth = half, twice). Only effects significant at $p=0.05$ are discussed here.

### 3.1 Formant bandwidth

Figure 2 shows the identification rate as a function of target and competitor bandwidth at each amplitude ratio. Dotted lines connect conditions that differ by the target's bandwidth. They all have a negative slope: all else being equal, identification was better for narrow- than for wide-formant targets. Full lines connect conditions that differ by the competitor's bandwidth. They also all have a negative slope: identification was better for vowels in competition with a wide- than a narrow-formant vowel. At -5 dB the lines form a parallelogram, indicating that the two effects are independent. At other ratios the shape is less regular, but this can be interpreted as the result of a sigmoid distortion reflecting ceiling and floor effects. Target and competitor bandwidth effects have similar sizes, with the result that identification is the same in the n/n and w/w conditions (except at -25 dB). A similar pattern prevailed at $\Delta F_0=6\%$ (not shown), with overall higher rates.

![Figure 2](image)

**Figure 2. Identification rate as a function of formant bandwidth (target/competitor), at each amplitude ratio, at $\Delta F_0=0$.**

Comparing effects of bandwidth at constant ratio with those of ratio at constant bandwidth (Fig. 2), it appears that a 4-fold change in bandwidth has an effect similar (in general slightly smaller) to that of a 10 dB change in ratio. Referring back to Fig. 1, a 4-fold reduction of bandwidth increases the formant peak amplitude by about 6 dB. One could thus explain target bandwidth effects by assuming (a) that narrowing a formant at constant RMS boosts its peak amplitude, and (b) that perceptual salience depends on the amplitude localized at formant peaks, rather than averaged over wider ranges, or the whole spectrum. The similar (and orthogonal) effects of competitor bandwidth suggest an analogous explanation: a vowel's masking power depends on the amplitude of its formants at their peak, rather than its RMS amplitude, or the spectrum level in the vicinity of the target's formants (as was hypothesized in the Introduction).

This explanation is problematic in at least two ways. First, it assumes precise sampling of the envelope amplitude at a formant peak, which is hard to reconcile with the smoothing steps that are often included in models of vowel perception. It also ignores the difficulty of estimating the amplitude of a narrow peak (45-55 Hz in the narrow condition) sparsely sampled by harmonics spaced at 124 or 132 Hz intervals [4]. Second, in the case of competitor bandwidth effects, the explanation supposes direct competition between vowels: the target is masked in proportion to the salience of the formant cues belonging to the competitor. This is in contradiction with conclusions of a previous study [5] that found evidence that cues to both vowels could coexist, as long as their own salience was not affected. This contradiction reveals limits of qualitative interpretations in terms of "feature salience", as performed in that study and here. Its resolution probably requires simulation with computational models of concurrent vowel perception.

![Figure 3](image)

**Figure 3. Improvement in target identification provided by a 6% $\Delta F_0$, as a function of target/competitor ratio, for each bandwidth condition.**

### 3.2 $\Delta F_0$ effects

Identification was better at $\Delta F_0=6\%$, as observed in many previous studies (eg. [1]). The improvement ($i(\Delta F_0=6\%) - i(\Delta F_0=0)$) is plotted in Fig. 3 as a function of amplitude ratio, for each bandwidth condition. The four lines have similar shapes. Their downward slope at large ratios reflects a ceiling effect: $\Delta F_0$ is of no help if identification is already perfect. Consistent with this idea, the right-hand edges of the four lines are staggered in inverse order relative to the rates plotted in Fig. 2 (In Fig. 3 $\Delta F_0$ effects extend to highest ratios for w/n, for which identification rates were lowest at...
ΔF₀=0 in Fig. 2). The drop-off on the left-hand side reflects the breakdown of segregation mechanisms at low target amplitudes.

It is interesting to note in this respect that ΔF₀ effects at -25 dB were larger for targets with formants that were narrow rather than wide. This might be an effect of the greater amplitude at the peak of their formants. Conversely, overall across ratios it appears that the magnitude of the ΔF₀ effect is larger for narrow- than for wide-formant competitors. It would seem that the masking power of a narrow-formant vowel surrenders more easily to a ΔF₀ difference.

3.3 Pairwise effects

Subjects’ responses may also be analyzed separately for each vowel pair. The appeal of such an analysis is that responses for each condition may then be compared with the spectrum of the stimulus for the same condition. The diversity of conditions (6 amplitude ratios) x (4 bandwidth combinations) for each pair allows a fine-grained analysis. The difficulty with this proposition is the large volume of data for the 20 target/competitor pairs, and the relatively small number of trials for each data point (30), that limits the reliability of effects observed [5]. Data for the pair /o+u/ will be presented in detail, to illustrate the potential of such an analysis, and also its limitations.

All single vowels had the same starting phase spectrum (Methods), and partials of same rank therefore summed in phase. There is thus no need to consider phase-dependent vector summation: the spectral envelope of a double vowel is simply the sum of the envelopes of its constituents, each
Formant bandwidth affects the identification of vowels in competition with other vowels. At constant RMS amplitude, identification of a vowel is enhanced by sharpening its formants, or widening those of its competitor. Effects of target and competitor bandwidth are approximately independent, and independent with those of amplitude ratio and $\Delta F_0$. The effect of a 4-fold change in target or competitor bandwidth is roughly the same order of that of a 6% $\Delta F_0$, or a 10 dB change in target/competitor amplitude ratio.

ACKNOWLEDGEMENTS

This work was performed within an agreement between ATR Human Information Processing Research Labs, CNRS, and University Paris 7. Rieko Kubo, of ATR ran the experiments, and Hideki Kawahara contributed valuable advice. The stimuli were synthesized with John Culling's |WAVE software.

REFERENCES


5. CONCLUSION

An unexpected outcome of the experiment was that sharpening a competitor’s formants increased, rather than decreased, its masking power, despite the fact that interformant valleys are deeper for narrow formant vowels. Target identification seems to depend on prominence of its formants relative to those of the competitor, rather than to the local spectrum level. Such a direct competition is inconsistent with data gathered in a previous study [5]. More research is required to resolve this question.

Formant bandwidth reflects losses within the vocal tract, for example damping during the open glottis phase. Such losses may vary with changes in phonation style, in particular stress. One could speculate that an effect of vocal stress might be to reduce acoustic losses (by shortening the glottal closed phase, and possibly stiffening tissues bounding the vocal tract), and thus give the speaker’s voice a competitive edge, enhancing both its masking power and its resistance to masking. If so, it might give the speaker an advantage in competitive social situations.