A PROBABILISTIC MODEL OF INTEGRATION OF ACOUSTIC CUES IN FV SYLLABLES

Fernández, S., Feijóo, S.

Departamento de Física Aplicada
Universidad de Santiago de Compostela
15706 Santiago de Compostela, SPAIN
E-mail: fasanti@usc.es

ABSTRACT

The interaction of consonantal and vocalic segments in FV syllables regarding identification of place of articulation of fricatives has been studied. A probabilistic model for integration of acoustic information in both segments is proposed. The model weights each segment's contribution and integrates them in order to resemble listeners' perception. First, the perceptual validity of the model has been assessed. Overall correlations of the probabilistic model with listeners' responses to a set of natural and conflicting-cue (place of articulation) FV syllables were 0.88 and 0.74, respectively. The results showed that in order to increase correlations, the model should weight F and V segments differently for each fricative, and even for each vocalic context, since listeners gave more or less importance to the vocalic transition depending on the particular fricative and vocalic context. Acoustic analysis was also carried out computing the probabilistic model with a posteriori probabilities for F and V segments obtained with quadratic discriminant analysis. The results were disappointing, indicating that usual statistical methods fail to extract coarticulatory information. This prevents ASR systems from taking advantage of the enhancement of the characteristics of the consonant in the FV condition with respect to the F condition.

1. INTRODUCTION

The purpose of this study is to research into the probabilistic nature of the consonant-vowel interaction in fricative-vowel syllables. Conflicting-cue stimuli are commonly used for studying the relative contributions of the acoustic cues contained in the consonant and in the vowel. Harris (1958) first studied that relationship in FV syllables, although her study was somewhat limited. Smits *et al.* (1996a) studied the integration of acoustic cues in stop-vowel utterances using both deleted-cue stimuli and conflicting-cue stimuli. They (Smits *et al.*, 1996b) tried to explicitly model the importance that listeners gave to both burst and transitions for the perception of place of articulation by combining a set of detailed and gross cues measured on the stimuli. Whalen (1984) also studied integration with conflicting-cue stimuli formed from both fricative-vowel and stop-vowel syllables. He found differences between both types of consonants, and concluded that listeners take cues into account even when those cues seem both superfluous and ineffective. This, therefore, makes difficult the selection of a set of acoustic cues in order to model integration.

Oden & Massaro (1978) proposed a theoretical probabilistic model for integration of speech features, but they were not particularly addressing the same topic as this paper. Neagu & Bailly (1995) tried to model integration using acoustic characteristics. They trained a classifier with the algorithm described in Nossair & Zahorian (1991) using natural stop-vowel syllables acoustically represented by a set of DCT coefficients that encoded the LPC spectrum of overlapping windows covering the whole syllable. The classifier was tested on conflicting-cue stimuli. Results poorly predicted listeners responses in the case of conflicting-cue stimuli. The importance given by the method to the consonant and the vowel did not agree with the listeners' perceptual evaluation of the tokens. In general, the method relies on the properties of the consonantal segment of the tokens in order to make a decision. The method performed far better for natural stimuli.

One of the most important effects of the aforementioned interaction is that perceptual recognition of fricatives when listeners are presented with both the fricative noise and the accompanying vowel is more accurate than when they are presented only with the noise cue. The magnitude of that effect depends on the particular fricative: it is big for /f, θ/, and very small or nonexistent for /s, ʃ/. Nevertheless, vocalic context also plays a role.

There are several possible explanations for that effect. One is to consider that the FV interaction is the result of a contrast effect: the quality of the vowel is reflected on some acoustic characteristics that are relational with respect to those of the fricative. Nevertheless, nobody has ever identified the acoustic features responsible for that effect. A second possibility is that coarticulatory processes impose some restrictions on the possible values of some acoustic characteristics of both the vowel and the fricative in a way that both the consonant and the vowel complement each other in the perceptual process.

In a previous experiment with conflicting-cue stimuli made up of FV syllables, both possibilities were evaluated from a perceptual point of view. The results showed that the second possibility may play a role for the non-sibilant fricatives. Vocalic context effects, however, also showed up, indicating that neither theory is able to completely explain the perceptual results.

In this paper a third possibility is considered. Since in both the F and V segments there is some kind of information relative to both the vowel and the consonant, it is possible that the auditory system evaluates that information in each segment, weights that information, and integrates them in one percept. Then, it is not necessary to have a very well defined acoustic characteristics in each segment to achieve a given percept, since two sources of information are available. A process of that kind would be more robust than one based on perfectly well defined cues.

2. MATERIALS

The stimuli were those of Fernández *et al.* (2000), where further details on the normalization and processing of the tokens can be seen. A male and a female native speakers of Spanish (Galician region) uttered a set of syllables formed by the voiceless fricatives /f, ɵ, s, ʃ/ followed by /a, e, i, o, u/. The tokens were sampled at 32 kHz. The friction noise (C^v) and the following 100 ms of the vowel (V^c) were excised in order to form the conflicting-cue stimuli:

- the 40 original tokens, C^v+V^c, were included as control stimuli.
- 160 C^v+V'^{c} tokens were used to study the importance of the theory of contrast in CV syllables.
- 120 C^v+$V^{c'}$ tokens served to assess the theory of integration.

Perceptual experiments with these stimuli were also carried out in Fernández *et al.* (2000). The original and contrast-based stimuli were equally well identified (93.7% and 90.0% correct, respectively) for the listeners, while integration-based stimuli attained a lower identification rate (74.0%), especially for /ɵ/ and /f/.

As regards the present acoustic analysis, the aforementioned stimuli form the test dataset. Tokens were resampled at 20 kHz. This reduces the order of the parameterization necessary to achieve a good acoustic characterization of the stimuli.

The training set was formed by the same syllables as the test set but pronounced twice by 15 male and 15 female speakers, digitized at 44.1 kHz and resampled at 20 kHz. Thus, 1200 tokens = 4 fricatives × 5 vowels × 15 speakers × 2 sexes × 2 repetitions were available for training and other 40 tokens = 4 fricatives × 5 vowels × 2 sexes formed the test dataset.

3. PROPOSED PROBABILISTIC MODEL

In order to propose a function modelling the relative contributions of the acoustic cues contained in the consonant and in the vowel, let us define $P_i(F)$ as the probability of membership of a fricative segment F in the class corresponding to one of the possible fricatives i = /f, ɵ, s, ʃ/, and $P_i(V)$ as the probability of membership of the transition cued by a vocalic segment V into the class corresponding to one of the possible fricatives i = /f, ɵ, s, ʃ/. Since the most perceptually relevant information for the identification of the fricative in FV syllables was found to be that of the noise along with the influence of the consonant in the following vowel (see Fernández *et al.* (2000)), the defined probabilities contain, in principle, the important information.

Our approach is also based on the hypothesis that the aforementioned probabilities, $P_i(F)$ and $P_i(V)$, can be determined from the F and V segments, respectively, as the a posteriori probabilities (APP) resulting from a classification procedure, say quadratic discriminant analysis (QDA), for example.

Under these conditions, the overall probability of a particular token corresponding to a particular fricative i can be estimated as:

$$P(i) = P_i(F) + P_i(V) - P_i(F) \cdot P_i(V) \qquad (1)$$

assuming that $P_i(F)$ and $P_i(V)$ are independent. This function represents the probability that a particular fricative i is determined by the fricative (F) segment or by the vocalic (V) segment. Obviously, this in only one of many possible choices of the probabilistic function. For example, a function stating that a particular fricative i is determined by the fricative (F) segment and by the vocalic (V) segment is simply:

$$P(i) = P_i(F) \cdot P_i(V) \qquad (2)$$

assuming again that $P_i(F)$ and $P_i(V)$ are independent, which, obviously, is not true for natural tokens. Nevertheless the assumption of independence is clearly fulfilled for conflicting-cue stimuli.

4. PERCEPTUAL VALIDITY OF THE PROBABILISTIC MODEL

In order to perceptually test the probabilistic model, we need to know the $P_i(F)$ and $P_i(V)$ probabilities assigned by the listeners to each token. Two perceptual experiments were carried out on the test dataset. Firstly, $P_i(F)$ were determined presenting the friction noise segment of the tokens to 37 listeners who were asked to identify them as one the possible fricatives /f, ɵ, s, ʃ/ or choose the option *another sound*. Therefore, $P_i(F)$ was the number of listeners that identified a particular fricative noise as the fricative i divided by the total number of listeners. Results of this perceptual experiment were commented in Feijóo *et al.* (1999).

Secondly, $P_i(V)$ were determined presenting the first 100 ms of the vowel to 25 listeners who were asked to identify them as one the following stops: /p, b, t, d/, or choose the option *vowel or diphthong*. The 25 listeners carried out the perceptual experiment three times. Results were pooled over listeners and repetitions. Responses that corresponded to the same place of articulation were merged into only one category: /p/ and /b/ responses were merged into *labial*, and /t/ and /d/ responses were merged into *dental*. Preliminary analysis of the results of this experiment showed that *labial* place of articulation responses were associated to /f/, *dental* responses were associated to both /ɵ/ and /s/, and *ambiguous* (*vowel or diphthong*) responses were associated to /ʃ/. $P_i(V)$ was the number of listeners that identified a particular vocalic segment as cueing place of articulation corresponding to fricative i, divided by the total number of listeners × 3 repetitions.

Once response profiles were available from both experiments, equation 1 was computed. The results of that equation should resemble listeners' responses to FV tokens. Response profiles for FV tokens corresponding to the natural syllables, contrast-based and integration-based conflicting-cue stimuli were also obtained in Fernández *et al.* (2000). Therefore they were correlated with the result of computing equation 1 with the probabilities of the corresponding F and V segments that formed each FV token. Table 1 shows those Pearson correlation coefficients for the natural, contrast-based and integration-based FV tokens, separately for each fricative along with the overall correlation including the four fricatives. Winner-takes-all (WTA) coefficients are also presented. Top table shows the results for equation 1. Equation 2 results are shown in the bottom table. As can be seen, correlation coefficients are quite high independently of the equation computed. WTA coefficients are very high indicating that the proba-

	natural		contrast		integration	
	r	WTA	r	WTA	r	WTA
f	0.88	1.0	0.87	0.95	0.67	0.87
θ	0.78	1.0	0.77	0.97	0.70	0.67
s	0.79	1.0	0.77	0.92	0.71	0.97
ʃ	0.94	1.0	0.93	0.97	0.92	0.90
overall	0.84	1.0	0.83	0.96	0.73	0.85
f	0.97	0.90	0.94	0.92	0.77	0.87
θ	0.89	1.0	0.86	0.97	0.75	0.67
s	0.91	1.0	0.87	0.92	0.74	0.90
ʃ	0.88	0.90	0.86	0.92	0.84	0.87
overall	0.88	0.95	0.84	0.94	0.74	0.82

Table 1: Pearson correlation and WTA coefficients for the natural, contrast-based and integration-based stimuli. Results are presented separately for each fricative along with the overall correlation including the four fricatives. Top table: results for equation 1; bottom table: results for equation 2. All correlations are significant at the $p < 0.0000005$ level.

bilistic model and the listeners coincided most of the times in the identification of the fricative presented in the FV syllable. Correlation coefficients are lower for the integration-based stimuli, especially when equation 1 is used and the perceptual identification of the FV syllable showed a greater effect of the transition (i.e. the F segment was /f/ , /θ/ or /s/ to a lesser extent). Equation 2 gives more weight to the vocalic transition than equation 1, and therefore the correlation for the integration-based stimuli is higher than when equation 1 is used as model. On the other hand, this implies that tokens including /ʃ/ obtain better correlations with equation 1, because for this fricative the friction noise is sufficient in order to perceptually identify the fricative.

Overall, the probabilistic model works quite well. Nevertheless, different fricatives need different weights for the noise and the transition. Thus a general model is not enough. Moreover, vocalic contextual effects were found in Fernández *et al.* (2000). A more detailed probabilistic model will imply the use of different weights for each fricative and vocalic context.

5. ACOUSTIC ANALYSIS

The acoustic analysis consists on the estimation of $P_i(F)$ and $P_i(V)$ from acoustic characteristics of the F and V segments, respectively. Both speech segments were covered with 15 ms wide windows overlapped 10 ms. In each window, 20 LPC-derived cepstral coefficients were computed. Three DCT coefficients encoded the trajectory of the 20 cepstral coefficients along the windows covering the speech segment. Thus 20 cepstral $\times$ 3 DCT coefficients = 60 variables were used to represent both the F segment and the V segment (Nossair & Zahorian, 1991).

Quadratic discriminant analyses (QDA) were carried out with the following classification groups for the fricative noise: /f/ , /θ/ , /s/ and /ʃ/. For the transition, we firstly tried, as categories for classification, the three places of articulation that listeners perceived: *labial* (for /f/), *dental* (for /θ/ and /s/) and *ambiguous* (for /ʃ/); nevertheless the QDA classified most of the /s/ tokens as *labial*, which does not agree with listeners' responses. The inability of the QDA to define the characteristics of the V segments with respect to the three same categories as listeners, led us to use the same four categories for the V segment as for the F segment. Since classification rates did not ever significantly improve increasing the number of included variables, the number of cepstral coefficients included in the QDA was always optimized in order to obtain the highest classification rate with the least possible number of coefficients.

The first QDA was performed on the training set. For the F segment, no significant improvement in the classification rate was obtained with more than 4 cepstral $\times$ 3 DCT coefficients = 12 variables. For the V segment 17 cepstral $\times$ 3 DCT coefficients = 51 variables were taken into account. The test dataset was then classified. Although the classification of the training set showed the expected pattern of confusions, the results for the test set were disappointing: for the F segment only 60% of the tokens were correctly classified and the pattern of confusions did not quite agree with the listeners, while 70% of the tokens were correctly classified from the V segment. Therefore a second QDA was carried out with the training and test datasets together using the leave-one-out method. In this case, for the F segment 11 cepstral $\times$ 3 DCT coefficients = 33 variables were taken into account, and 20 cepstral $\times$ 3 DCT coefficients = 60 variables for the V segment. Both the classification rates for the fricative noise segments (84% for the tokens of the test dataset) and the pattern of confusions were according to listeners' perception (overall correlation was 0.76, 0.81 for /f/, 0.66 for /θ/, 0.72 for /s/ and 0.93 for /ʃ/). For the vocalic segment only 27.5% of the tokens of the test dataset were correctly identified, although the overall classification rate with the training and test datasets together was 72.6%. The correlation of the test tokens' probabilities with listeners' responses was not significant in any case. Tokens with /s/ transitions were mostly identified as tokens with /θ/ transitions, which is correct because both are dental transitions, and tokens with /θ/ transitions were mostly identified as tokens with /s/ transitions. /ʃ/ transitions were clearly different from the rest of transitions. /f/ transitions were mostly identified as /θ/ transitions.

Obviously, our hypothesis stating that the probabilities $P_i(F)$ and $P_i(V)$ can be determined from the F and V segments, respectively, as the a posteriori probabilities (APP) resulting from a classification procedure like QDA is not being confirmed by the acoustic analysis, which has serious difficulties, especially extracting the information relevant to the transition of a particular fricative from the V segment. There are more evidences about this fact yet. Equations 1 and 2 were computed using the aforementioned APP's. The acoustic probabilistic profiles obtained were correlated with the listeners' response profiles in the FV condition. Table 2 shows the pearson correlation coefficients for the natural, contrast-based and integration-based stimuli. Correlation coefficients for the probabilistic model based on QDA on a training and a test dataset are low (approximately 0.5) when equation 1 is applied. Computing equation 2 even lowers the correlation with respect to that for equation 1, indicating that information in the V segment relevant to the transition is not well represented by the probabilistic profile $P_i(V)$. Using a single dataset and the leave-one-out method in the QDA does not help very much. Correlations are slightly higher (around 0.6) but also now they are lower when equation 2 is used than when equation 1 is computed.

	QDA				Neural Network	
	train & test		leave-one-out			
	Equation 1	Equation 2	Equation 1	Equation 2	Equation 1	Equation 2
natural	0.54	0.36	0.65	0.57	0.53	0.53
contrast-based	0.53	0.40	0.63	0.59	0.54	0.57
integration-based	0.50	0.31	0.27	0.15*	0.25	0.09**

Table 2: Pearson correlation for the natural, contrast-based and integration-based stimuli. Results are presented separately for each classification method and proposed probabilistic model, i.e. for equation 1 and for equation 2. All correlations are significant at the $p < 0.0000005$ level, except * that was significant at the $p < 0.001$ level, and ** that was not significant.

They are especially low (around 0.2) for the integration-based stimuli, for which the fricative information in the V segment is particularly important. Again, the probabilistic method fails in those conditions in which the conflicting-cue is the place of articulation, and therefore a good characterization of the vocalic transition is fundamental. While the method seems to be quite adequate for extracting the fricative information in the noise, it is unable to obtain from the vocalic transition the relevant characteristics for the identification of fricative place of articulation.

Another striking point showed up. The probabilistic nature of the statistical method was in doubt, because the probabilities assigned by the analysis to each category were extreme, i.e. one or nothing, this effect being independent of whether the token is assigned to the *correct* class of not. On the other hand, listeners' judgments tend to result into more *fuzzy* response profiles, and this is what really validates a probabilistic model like the one that is proposed. In order to assess what was the importance of this effect a conventional, fully connected, feed-forward Neural Network (NN) with three layers was used in order to determine the probabilities $P_i(F)$ and $P_i(V)$. Neural networks are supposed to generalize better than classic statistical methods and APP's can be estimated accurately. The results showed that tokens were represented probabilistically, i.e. the values of the probabilities were not so extreme. Nevertheless, the results were very similar to those obtained with QDA (see table 2), confirming that difficulties for characterizing the transition of a particular fricative are the most serious obstacle that has to be overcome.

6. CONCLUSION

The proposed probabilistic model integrating information from both the F and V segments in FV syllables works acceptably well identifying fricatives. These results suggest that listeners weight the information provided by each segment and integrate them in one percept. Although a general probabilistic model works well given the same weight to each segment in every case, it has been proved that a detailed model of integration of this kind needs to employ different weights for both segments depending on the fricative, and even depending on the vocalic context, in order to perfectly match listeners' perceptual identification. This is a consequence derived from the fact that the perceptual identification of different fricatives in FV syllables relies to a different extent on the vocalic transition, this effect depending on the vocalic context.

As regards the acoustic modelling of that interaction and the possibility that ASR systems take advantage of them, this paper shows that the acoustic characterization of the vocalic transition is still far from being satisfactory to provide relevant information for the identification of place of articulation of fricatives. The probabilistic model failed in those conditions in which the listeners gave more importance to the vocalic transition. This questions the efficiency of usual statistical methods for extracting coarticulatory information. Three possibilities arise: the acoustic variables employed are unadequate; the training dataset is not big enough; the statistical model is unable to group the relevant characteristics. It is quite possible that the combination of those three points is responsible for the lack of success.

7. REFERENCES

1. Feijóo, S., Fernández, S., Balsa, R., "Influence of frequency range in the perceptual recognition of fricatives," *J. Acoust. Soc. Am.*, vol. 105, No. 2, p. 1401(A), 1999.
2. Fernández, S., Feijóo, S., Balsa, R., Barros, N., "Perceptual effects of coarticulation in fricatives," *Proceedings of ICASSP 2000*, vol. III, pp. 1347–1350, 2000.
3. Harris, K. S., "Cues for the discrimination of American English fricatives in spoken syllables," *Lang. and Speech*, vol. 1, pp. 1–7, 1958.
4. Neagu, A., Bailly, G., "Relative contributions of noise burst and vocalic transitions to the perceptual identification of stop consonants," *Proceedings of Eurospeech'95*, 1995.
5. Nossair, Z. B., Zahorian, S. A., "Dynamic spectral shape features as acoustic correlates for initial stop consonants," *J. Acoust. Soc. Am.*, vol. 89, pp. 2978–2991, 1991.
6. Oden, G. C., Massaro, D. W., "Integration of featural information in speech perception," *Psychological Review*, vol. 85, pp. 172–191, 1978.
7. Smits, R., ten Bosch, L., Collier, R., "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment," *J. Acoust. Soc. Am.*, vol. 100, pp. 3852–3864, 1996a.
8. Smits, R., ten Bosch, L., Collier, R., "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modeling and evaluation," *J. Acoust. Soc. Am.*, vol. 100, pp. 3865–3881, 1996b.
9. Whalen, D. H., "Subcategorical phonetic mismatches slow phonetic judgments," *Percept. Psychophys.*, vol. 35, pp. 49–64, 1984.