

Comparing HMM-based Recognition to Perceptual Phonetic Integration in Fricative-Vowel Syllables

Santiago Fernández^{*†} and Sergio Feijóo[†]

^{*} Department of Phonetics and Linguistics, University College London, UK
santi@phon.ucl.ac.uk

[†] Departamento de Física Aplicada, Universidad de Santiago de Compostela, SPAIN

ABSTRACT

In fricative-vowel syllables, the perceptual identification of the consonant is significantly better when at least part of the accompanying vowel is available, with respect to the case in which only the consonant is heard. In this study, the ability of hidden markov models (HMM) for reproducing the process of phonetic integration is assessed by comparing automatic recognition with listeners' perception. Results show that HMM perform better than listeners for isolated fricative noises, but worse for fricative+vowel segments. The utility of fricative+vowel models versus syllable models is discussed.

1 INTRODUCTION

The perceptual mechanism by which the components of a syllable interact to enhance the identity of the constituent phonemes is still unknown [1, 2, 3]. Due to this process of phonetic integration, for fricative-vowel syllables, the perceptual identification of the consonant is significantly better when the accompanying vowel is available with respect to the case in which only the consonant is heard.

Current ASR systems do not take full advantage of phonetic integration mechanisms. They are based on either monophoneme models or treat the interaction among phonemes as distortion that must be accounted for using specific models for each phonetic context. Another alternative is the use of larger speech units, such as syllables, but this implies a high number of complex models and loss of flexibility. Thus, large databases are needed to train those models. Therefore, this approach has not yet been fully exploited in continuous speech recognition tasks.

It is not clear yet if listeners need to identify the whole syllable or if they need to identify every component to make use of phonetic integration mechanisms. Including only part of the surrounding phonemes can be enough to improve perceptual identification of the tar-

get phoneme. This helps to take contextual information into account while maintaining a small number of classification categories.

In this paper, we assess the ability of hidden markov models (HMM) for reproducing the process of phonetic integration by comparing their performance with listeners' perception of isolated fricative noises and fricative noises plus part of the following vowel.

2 MATERIALS

Fricative-vowel syllables were formed by the combination of fricatives /f, θ, s, ʃ/ and /x/ with vowels /a, e, i, o/ and /u/. Syllables were pronounced by 39 male and 37 female native speakers of Spanish (Galician region). Signals were sampled at 32 kHz with a precision of 16 bits, and then high-pass filtered at 100 Hz in order to avoid any undesired breathing noise. After that, they were normalized so as to make use of 75% of the quantization range. Finally, the signals were segmented in order to obtain two different types of tokens: isolated fricative noise (henceforth F condition), and isolated fricative noise plus 100 ms of the following vowel (henceforth FV condition).

Tokens were separated into two datasets: a test dataset and a training dataset. The test dataset is presented to listeners and it is also used to assess the validity of the acoustic models developed with the training dataset. For each condition (F and FV), in the training dataset there is a total of 2800 tokens = 5 fricatives × 5 vowels × (29 male + 27 female speakers) × 2 repetitions; and in the test dataset there are 500 tokens = 5 fricatives × 5 vowels × (10 male + 10 female speakers).

3 PERCEPTUAL IDENTIFICATION

Perceptual identification of the test dataset indicates what should be the performance of the classification analysis. A detailed analysis of the results of the perceptual experiment has already been reported [3]. We

	/f/	/θ/	/s/	/ʃ/	/x/	other
/f/	63.3	21.2	2.3	3.8	3.2	6.2
/θ/	45.2	36.4	4.0	4.0	4.5	5.9
/s/	2.4	10.1	55.9	30.0	0.7	0.9
/ʃ/	1.2	5.2	18.9	73.9	0.1	0.7
/x/	2.6	2.2	0.8	2.7	87.2	4.5
/f/	85.1	13.4	0.3	0.4	0.8	0.0
/θ/	14.5	83.0	1.0	0.2	1.2	0.1
/s/	0.3	6.9	76.3	16.3	0.1	0.1
/ʃ/	0.4	1.1	8.8	89.4	0.1	0.2
/x/	2.0	0.7	0.2	0.2	96.2	0.7

Table 1: Listeners’ Confusion matrices. Top: F condition (63.3%); bottom: FV condition (86.0%).

will give here only a summary of those results.

3.1 METHOD

Perceptual experiments were carried out on the test dataset. Twenty one subjects participated in the experiments for course credits. The tokens were presented blocked by condition. Listeners had to identify the fricative, firstly in the F condition, then in the FV condition. They could hear each stimulus up to four times, then they had to select one of the following responses: /f, θ, s, ʃ, x/ or *other*. This last option was available just to verify that isolated fricative noises had been identified as speech sounds.

3.2 RESULTS

Confusion matrices for both conditions are shown in Table 1. Fricatives are better identified in the FV condition than in the F condition (86% versus 63%). The difference between conditions is statistically significant ($F(1, 998)=574.26, p<0.0005$), even for each fricative separately. As can be seen, confusions take place between the pairs /f-θ/ and /s-ʃ/, indicating the acoustic similarity between phonemes in both pairs. Fricative /x/ is well identified in both conditions.

Inter-listener consistency for each condition will give us an objective measure both of the ambiguity of the tokens in the dataset and of the highest attainable correlation between listeners’ identification and automatic recognition. Inter-listener consistency has been computed using the *split-half* method which consists in dividing the sample of listeners in two groups and computing the correlation between the responses given by both groups. Correlation was $r=0.91$ in the F condition and $r=0.97$ in the FV condition. This indicates that the FV condition is perceptually less ambiguous than the F condition.

There are, nevertheless, substantial differences among fricatives. In the F condition inter-listener consistency was 0.91, 0.68, 0.89, 0.92, 0.98 and 0.35 for /f, θ, s, ʃ, x/ and *other*, respectively. Results for the FV condi-

Token No.	/f/	/θ/	/s/	/ʃ/	/x/	other
1	12	7	0	0	1	1
⋮			⋮			
n	1	0	0	0	20	0

Table 2: Example of a response profile.

tion were 0.97, 0.96, 0.96, 0.97, 0.99 and 0.73. Inter-listener consistency for option *other* indicates that in the F condition, the majority of *other* responses are due to the natural ambiguity of the isolated fricative noise, while in the FV condition the majority of *other* responses are due to particular stimuli that do not have well defined characteristics (e.g. because of breathing noise). These figures indicate that stimuli were perceived as speech sounds and therefore validate the results of the perceptual experiment.

4 CLASSIFICATION ANALYSIS

The methodology employed in the acoustic analysis has been chosen with the purpose of allowing the comparison, to some extent, between previous analysis on the same dataset [4] and the new analysis reported in this paper.

4.1 METHOD

Signals were resampled at 20 kHz in order to reduce the number of parameters in the model while maintaining the acoustic characteristics which are relevant for the perceptual identification of fricatives. All other aspects of the processing of the signals are identical to those reported for the perceptual experiments.

Both the acoustic characterization and classification analysis was carried out with the HTK Hidden Markov Model Toolkit (<http://htk.eng.cam.ac.uk/>). Stimuli were acoustically characterized by a set of twenty MFCC computed over consecutive windows of 15 ms, with an overlap of 10 ms, which covered the entire segment corresponding to the F or FV condition. The number of filterbank channels was 27. In the F condition, three state left-to-right hidden markov models were used for each one of the five fricatives. In the FV condition, both three and five state models were tested. Output probabilities were modelled by a mixture of Gaussians. Models with four and eight gaussians were tested. Firstly, the training dataset is used to estimate the parameters of the models and, secondly, the test dataset is evaluated using the trained models.

The results for the test dataset can be compared with listeners’ identification. Results of perceptual experiments have been stored as *response profiles*: for each stimulus, the total number of responses given to each option is stored (see Table 2). This gives us a measure

of the distance (actually an *inverse distance*) between each stimulus and each phonetic category. Similarly, *a posteriori* probabilities (APP) of membership into each phonetic category, obtained from a statistical analysis, give us an objective measure of the distance between each stimulus and each phonetic category. This is called a *classification profile*.

Perceptual identification and automatic recognition can be compared efficiently by computing the correlation between response and classification profiles. A *class* correlation can be obtained by computing the correlation coefficient between vectors of the response and classification profiles associated to the same phonetic category. An *overall* correlation can be obtained by computing the correlation between vectors obtained by concatenation of all *class* vectors. Before computing Pearson correlation, all vectors have been transformed to have zero mean [5].

Unlike other statistical methodologies, HMM lack discriminative power and acoustic scores are obtained as likelihoods instead of as APP. Fortunately, Bayes' rule allows us to compute APP from likelihoods:

$$P(M|O) = \frac{P(O|M) \cdot P(M)}{P(O)} \quad (1)$$

where $P(M|O)$ is the *a posteriori* probability of model M being recognised given the acoustic observation O ; $P(O|M)$ is the likelihood obtained from HMM (probability of generating observation O with model M); $P(M)$ is the *a priori* probability for each model, which is, in our case, constant and equal to one divided by the number of phonetic categories; and $P(O)$ is the *a priori* probability of observation O , which is equal to:

$$P(O) = \sum_M P(O|M) \cdot P(M) \quad (2)$$

Using both equations and taking into account that $P(M)$ is constant:

$$P(M|O) = \frac{P(O|M)}{\sum_M P(O|M)} \quad (3)$$

4.2 RESULTS

Classification matrices corresponding to the highest correlation coefficients can be seen in Table 3 for both conditions. Recognition rates for the F condition are higher than listeners' identification rates (80.4% versus 63.3%). There are only slight differences between recognition rates in the F and FV conditions (80.4% versus 81.2%). In the FV condition, listeners' identification rates are higher than those obtained by the acoustic analysis (86.0% versus 81.2%). Confusions take place between /f-θ/ and between /s-f/ as well. Nevertheless, there are some differences with respect to listeners' identification. Fricative /θ/ is well recognised even with the fricative noise alone. More /f/ fricatives

	/f/	/θ/	/s/	/f/	/x/
/f/	68	28	0	3	1
/θ/	16	75	7	0	2
/s/	0	7	89	4	0
/f/	0	0	27	73	0
/x/	0	0	0	3	97
<hr/>					
/f/	73	20	5	1	1
/θ/	15	75	9	0	1
/s/	0	3	92	5	0
/f/	0	0	25	75	0
/x/	1	2	3	3	91

Table 3: Classification matrices. Top: F condition (80.4%); bottom: FV condition (81.2%).

are recognised as /s/ than the other way round. The reverse is true for listeners.

The best overall correlation coefficient in the F condition was obtained with three states and four gaussians ($r=0.77$)¹. In the FV condition the best overall correlation was obtained with five states and eight gaussians, $r=0.81$, which is only slightly higher than the correlation in the F condition.

This indicates that the automatic classification does not benefit from the inclusion of the vowel to the same extent as listeners do. Actually, comparing these correlations with inter-listener consistency (0.91 and 0.97 in the F and FV conditions, respectively) it is clear that the acoustic analysis can be further improved. *Class* correlations can be of some help for this task. In the F condition, the correlation between response and classification profiles were 0.69, 0.70, 0.84, 0.78 and 0.96 for /f, θ, s, f/ and /x/, respectively. This means that the acoustic characterization of the /s-f/ pair and, especially, of the /f-θ/ pair should be improved. In the FV condition correlation coefficients were 0.80, 0.74, 0.78, 0.80, 0.95, respectively; indicating that, at least for the /f-θ/ pair, the acoustic characterization can be improved by including the vowel [6].

While recognition rates are clearly higher than listeners' identification rates in the F condition, HMM do not benefit from including the vowel to the same extent as listeners do. One of the possibilities to explain this, is that HMM work well when the variance in the sample is limited to that of the classification categories. New information, such as the vowel in our case, which is not exclusively related to the classification categories, implies more variance. If there is a balance between the new information provided and the extra-variance included, HMM might obtain similar results in both the FV and in the F condition; otherwise, performance can be either degraded (too much extra-variance) or improved (good new information provided).

¹All the correlation coefficients reported are highly significant ($p < 0.0000005$).

To clarify this point to some extent, HMM were trained, in the FV condition, using the syllabic category of the stimuli. This approach clearly reduces the variance in the sample at the expense of increasing the number of models from five fricatives to twenty-five syllables. If variance is the reason why HMM did not benefit from including the vowel as expected, this new approach should give better results. In order to compare this new results with listeners perception, we need a classification profile with phonetic categories (fricatives) instead of syllabic categories. To obtain the APP of membership of a particular stimulus into a particular fricative class, we sum the APPs of membership of that stimulus into each of the syllabic categories containing the fricative. Then, correlation coefficients can be computed.

The best overall correlation coefficient with syllables was obtained with five states and four gaussians ($r=0.84$). Class correlations were 0.81, 0.78, 0.84, 0.81, 0.97 for /f, θ, s, ʃ/ and /x/, respectively. These figures are only slightly higher than those obtained in the FV condition. The confusion matrix is very similar to that obtained in the FV condition (already shown in Table 3). Percent correct for fricatives was 83% versus 81.2% in the FV condition.

To sum up, overall classification rate in the F condition is almost 20% higher than that obtained by listeners, while in the FV condition overall classification rate is 5% lower than that obtained by listeners. Correlation coefficients for both conditions are still far from inter-listener consistency in both conditions, indicating that acoustic analysis can still be improved. HMM do not benefit from including the vowel to the same extent as listeners do. For the /f-θ/ pair results are better when the vowel is added. There is not much difference between overall performance of both approaches considered in this paper: adding vocalic information (FV condition) and adding vocalic information plus reducing variance to some extent (syllable condition).

5 DISCUSSION

HMM perform better in the F condition than other discriminative approaches such as discriminant analysis and neural networks [3, 4] for the tokens in this study. In the FV condition results are similar. One striking outcome is that overall classification rate in the F condition is 20% higher than that obtained by listeners, indicating that there is a large difference between the acoustic approach and perceptual mechanisms.

Adding (at least part of) the vowel helps listeners to identify clearly the fricative. HMM do not benefit, to the same extent as listeners, from the new information available when the vowel is added. Nevertheless, adding the vowel is useful for particular fricatives such

as /f-θ/ and, in some conditions, such as noisy speech, including the vowel might improve performance to a further extent.

This inability of HMM does not seem to be due to the presence of new variance (related to the acoustics of the vowel) that cannot be accounted for, since results are similar for the FV and syllable conditions. When the inclusion of the vowel is important, FV models might be useful with small, and maybe acoustically compact, training databases, since the number of models is drastically reduced. This approach might require a segmentation stage, previous to the training of the FV models, that could be carried out with monophoneme models. Syllable models seem useful for large training databases with great variety of speakers and forms of speech.

ACKNOWLEDGMENTS

Part of this research has been supported by a Marie Curie Fellowship of the European Community programme "Improving the Human Research Potential and the Socio-Economic Knowledge Base" under contract number HPMF-CT-2002-02129.

Many thanks to the members of the Grupo de Tratamiento de la Señal, Universidad de Vigo, for their help with the acoustic analysis.

REFERENCES

- [1] K. Kurowski and S. E. Blumstein, "Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants," *J. Acoust. Soc. Am.*, vol. 76, pp. 391–404, 1984.
- [2] D. H. Whalen, "Three lines of evidence for direct links between production and perception in speech," in *Proc. ICPHS*, San Francisco, 1999.
- [3] S. Fernández, *Integración fonética en sílabas fricativa-vocal*, Doctoral Thesis, Universidad de Santiago de Compostela, 2001.
- [4] S. Fernández and S. Feijóo, "Computational models of integration in FV syllables," in *Proc. of ISCA Workshop on TIPS*, Aix en Provence, 2002.
- [5] J. S. Bendat and A. G. Piersol, *Random data: Analysis and measurement procedures*, New York: Wiley-Interscience, 1971.
- [6] K. S. Harris, "Cues for the discrimination of American English fricatives in spoken syllables," *Lang. Speech*, vol. 1, pp. 1–7, 1958.