DOI: 10.61782/fa.2023.0529

A MACHINE LEARNING MODEL TO ASSESS THE INTEGRATION OF VISUAL AND AUDITORY CUES DURING SPEECH PERCEPTION IN NOISE

Mark Gibson[1*] **Mónica González Machorro**[2] **Marcel Schlechtweg**[3]

[1] Department of Linguistics, Speech Laboratory, Universidad de Navarra, Spain
[2] Digital Health — Connected Healthcare, Hasso Plattner Institute, Universität Potsdam, Germany
[3] School of Linguistics and Cultural Studies, Carl von Ossietzky Universität Oldenburg, Germany

ABSTRACT

We trained a set of K-means clustering models in an unsupervised learning environment on acoustic and visual data to address how the visual and auditory modalities interact when accessing phonological categories in the face of a noise source (which introduces entropy in the flow of information). Our objectives for the models were to explain the results of a previous round of psychoacoustic perception experiments using auditory stimuli only, and to make predictions as to how including a visual stimulus, in this case lip aperture (vocal tract variable), would affect response accuracy across different groups, which we will corroborate in future psychoacoustic/visual tests. We discuss the import of our findings in relation to audio-visual integration during speech perception, and map future paths to address this issue in different populations.

Keywords: *noise, vowel discrimination, auditory cues, visual cues, machine learning model, speech perception.*

1. INTRODUCTION

Crucial to Shannon's [1] mathematical model of communication is the idea that a noise source acts on some signal, which subsequently challenges the flow of information encoded in that signal. As much holds for computationally derived messages as for human speech. Essentially, an information source encodes a message that a subsequent transmitter converts to a signal by way of some channel, in our case an acoustic channel. The receiver unpacks the information encoded in the channel(s) in order to decode the message intended by the transmitter. Noise, at the same time, is intrinsically related to entropy (uncertainty or randomness) in that noise degrades the information content contained in the message.

Gibson et al [2] examined linguistic effects and interactions when discriminating vowels in two noise conditions (computer generated background babble and modifications to the signal-to-noise ratio, henceforth SNR). The number of background speakers was set from 1-16. Native English- and Spanish-speaking adults were administered the same test in which the syllables 'da, de, di, do, du' appeared on a screen while an auditory stimulus of one of those syllables was played in the midst of background babble and modifications to the SNR, which were applied randomly by the MATLAB-based test. Participants were instructed to click over the box containing the syllable they heard. Stimuli were presented for two speakers, one biological female, and one biological male. All routine normalizing was performed on the stimuli (duration, volume, etc.), and all tests were administered according to standard perception-testing protocol (i.e. using professional sound equipment, sound-proof booth, quiet environment etc.).

Schlechtweg et al (unpublished data as of yet) employed a similar test modified for children with typical hearing, henceforth TH, and their cohorts with cochlear implants, henceforth CI, the only difference being that the SNRs were set to 0, 6 and 12 dB and the number of background speakers was programmed to a constant 6, in order to shorten the test and ensure maximum attention.

In a tangential study (unpublished data as of yet) measured the development of masking release (the capacity to differentiate, or release, a target stimuli from its masker) as a function of age in typical-hearing children ages 6-12 (N=35), again using the modified version of the original adult version of the test in order to ensure maximal attention.

[*] ***Corresponding author**: mgibson@unav.es*

Schlechtweg et al (mentioned previously) found that when considering all vowels together, the children with TH (77.0 %) responded, on average, more accurately than the CI group (72.6 %), though only slightly. A by-vowel analysis revealed a relatively high response accuracy for [a], [e] and [i] for both groups. Interestingly, the children with cochlear implants answered more accurately than the children with typical hearing for [a] and [e]. Discrimination of the high-back vowel [u], however, proved more taxing for the children with CI, who responded less accurately to this vowel than (a) they did to all other vowels, and (b) than did the children with TH. The children with CI had observably lower scores than the TH group in detecting [u] in noise. Further, the CI group confused [u] with [o] in ± 45 % of the cases in which [u] was presented as the stimulus (these results mirror the results in [2] and [4]). As the authors in these studies surmise, this confusion of [u] with [o] may be a consequence of the fact that [o] and [u] are back vowels and therefore have close second formant (F2) values [see, e.g., [4, 5]]. Previous studies have shown, for example, that individuals with hearing loss tend to confuse vowels with the vowel closest to them in terms of first formant (F1) or F2 values. Confusion of [o] and [u] has also been shown to occur empirically in other languages such as Dutch [see [6]], in addition their existing historical alterations in other Romance languages such as Catalan, Romanian and Portuguese, as well as Northern Peninsular dialects of Spanish ([7-9]]. These findings fit nicely into the cluster idea [10], in that when confronted with high overlapping inter-categorical variation and high intra-categorical overlap, a vowel from one category of a cluster can often be interchanged with a vowel from an overlapping contiguous category (e.g., [o] and [u] overlap in that both are back rounded vowels, albeit with differing qualifications in the vertical dimension). This concept is buttressed by the finding that vowel categories tend to overlap to a greater extent for listeners with hearing impairment [10]. To account for this, Gibson et al [2 and 3] postulate that the third formant (F3) value, which correlates physically to lip rounding, and is similar for [o] and [u], may be masking the F1 in these cases, obfuscating perceptual access to vowel height [see, e.g., [4]].

As for the effects of the specific noise types, Schlechtweg and colleagues (unpublished data as of yet) notes that for the TH group, the [o]-[u] confusion did not represent a dominant pattern, either overall, or for any specific SNR parameter. While [o] for [u] appeared in 5 % of the cases for the 0 dB SNR, they did so as well in 5% and 8.8 % of the cases at the 6 and 12 dB SNRs, respectively (the response accuracy for [u] was 60 % (0 dB), 83.8 % (6dB), and 78.8 dB (12 dB)) (these figures are also in line with the results presented in [2 and 3]). For the CI group, however, in almost half of the trials the subjects chose [o] for [u] for all SNRs. Moreover, clear differences were observed across the individual SNRs. While [o] was selected in 23.8 % of the cases where [u] was played for the 0 dB condition, it was chosen in 51.2% and 58.8 % of the trials at 6 and 12 dB, respectively (the response accuracy for [u] was 31.2 % (0 dB), 45 % (6 dB), and 38.8 dB (12 dB)). That is, even when there is clearly more signal than noise, the children with CI exhibited less accuracy in detecting [u], and primarily confused [u] with [o].

For [2] and [3] results were commensurate with the general findings showing confusion in discriminating the rounded back vowels [o] and [u] in noise. It is noteworthy that at 0 dB SNR there was nearly perfect vowel discrimination but any errors at 0 dB SNR were almost all due to [o,u] confusions. To explain the catholic back rounded vowel confusion across all subject groups tested, for the current study we performed a detailed acoustical analysis of the stimuli (produced by both speakers whom we recorded for the stimuli, one biological male/one biological female) in order to examine whether there was a physical impetus for the misperception. Our analysis showed greater acoustical distance, with regard to F1, between the back vowels for the biological female speaker than for the biological male, who had very close F1 values for [o] and [u] (though not overlapping). However, the biological female speaker showed closer values for F1 for the front vowels, meaning that if acoustic density alone were the motivation for the confusion between [o] and [u], the confusion effect would be stronger for the front vowels than for the back vowels, which does not find support in the data. This observation also lends evidence to the idea presented in [2-4], that F3 may be masking F1, since confusion of the back rounded vowels is still relatively high even for the biological female stimuli even though there is more acoustic differentiation in F1 for the back rounded vowels than for the front vowels. Mean values (in Hz) for F1, F2, F3 and F0 (fundamental frequency) are given in Table 1 for each speaker.

Table 1. Mean formant values (rounded to conserve space, in Hz) by speaker (where F=Female and M=Male, both biological) for the stimuli vowels [a,e,i,o,u] in Spanish /dV/ syllables.

Speaker	Vowel	F1	F2	F3	F0
F	[a]	973	1899	2826	230
	[e]	473	2474	3046	219
	[i]	421	2801	3203	223
	[o]	569	1381	2964	290
	[u]	428	1161	2939	230
M	[a]	692	1498	2633	118
	[e]	419	1994	2521	115
	[i]	260	2305	3145	120
	[o]	413	1108	2696	150
	[u]	413	1322	2854	117

F1/F2 (where '/' means '÷') and F1/F3 ratios were also calculated given their importance in vowel discrimination (and are offered in Table 2). In [12], for example, the authors showed magnetoencephalographic evidence that the auditory cortex is sensitive to formant ratios and plays a role in vowel discrimination.

Table 2. F1/F3, and F2/F3 ratios for all stimuli vowels by speaker.

Speaker	Vowel	F1/F3	F2/F3
F	[a]	0.34	0.67
	[e]	0.16	0.81
	[i]	0.13	0.87
	[o]	0.19	0.47
	[u]	0.16	0.81
M	[a]	0.26	0.57
	[e]	0.17	0.79
	[i]	0.10	0.73
	[o]	0.15	0.41
	[u]	0.16	0.81

As can be observed in Table 2, there is only a scant difference for the F1/F3 ratio for [o] and [u] (15% and 16% respectively) for the biological male speaker, while the biological female exhibits greater differentiation for the F1/F3 ratio for [o] and [u] (19% and 16% respectively).

To account for the general confusion of the back rounded vowels across a wide range of subjects, [2 and 3] surmised that in absence of a visual input, such as lip aperture or jaw angle, which distinguishes [o] and [u] (and the other vowel categories), and would otherwise serve to reduce uncertainty or entropy, the subjects lack crucial information (information flow is essential to Shannon's 1948 model) to which they would else be privy in non-laboratory-based settings, that may aid in discerning the phonological contrasts for the different vowel categories. Against this backdrop, in what follows, we present the results of a set of machine-learning models in an unsupervised environment based on K-means clustering in order to address the research question as to how visual information and acoustical information interact to better predict phonological contrasts across a wide range of subjects. The parameters and model design are reported in full in the sections that follow.

2. CLUSTERING MODELS

2.1 K-means clustering

K-means clustering is an unsupervised method that clusters unlabeled data. It is a method of vector quantization and is one of the most popular unsupervised methods in data mining. It requires a fixed number of clusters (k) and features. Based on the input, the algorithm iteratively calculates the positions of the centroids while keeping them as small as possible. The goal of this algorithm is to minimize the sum of the squared Euclidean distances of each point to the closest cluster. To implement our models we used scikit-learn's [13] K-Means class in Python.

2.1.1 Model parameters and evaluation metrics

The use of an unsupervised environment is grounded in the intuition that a learner (say an infant learning vowel categories) has no *a priori* reason to assume any underlying categories (hence they learn categories from the data). The first model was built on randomly generated data (N=200,000; 5 vowels x 20,000 trials x 2 speakers) using the productions from the speakers as the baseline whom we recorded for the stimuli for the psychoacoustic tests presented in Section 1 [2 and 3]. We used F1, F2, F3 and F0 as the variables most likely to distinguish vowel categories (though, we do not discard using ratios F1/F3 and F2/F3 and modulations of envelope amplitude, ΔE, for future models). Random values were computed using one standard deviation point above and below the mean, which is a conservative measure given the nature of formant value variation across contexts. The second model was trained on the acoustic data in addition to being fitted with visual data, maximum lip aperture to be specific. Visual data were collected for both speakers whose data served for the acoustical model. Visual recordings (videos) were made using personal mobile devices and converted to mp4 files. The mp4 files were converted to motion capture using Python, as shown in Figure 1 (left), where each image along the x-axis represents an image at every ± 2 ms timestamp in the video file (motion capture frames were generated at an interval of 30 frames per second). We then traced the dynamic movement across frames in order to find maximum lip aperture. Pellets were placed on the upper and lower lips to facilitate the collection of quantitative measures (in mm) in ImageJ, as shown in Figure 1 (right). In order to perform these measures, pixels were scaled to physical measures by registering images of a known distance (using a ruler) to the number of pixels in the same image. We measured from the bottom-most edge of the superior lip pellet to the top-most edge of the inferior lip pellet. We report the values in mm in the following Table 3.

To add noise to the models, we randomly added white noise as an additional variable to the acoustic data by using Normal Gaussian distribution, where:

$$\mu = 0 \text{ and } \sigma = spectral\,noise\,density\,unit\ (Hz).$$

For our experiments, we defined the spectral noise density unity as 1 Hz.

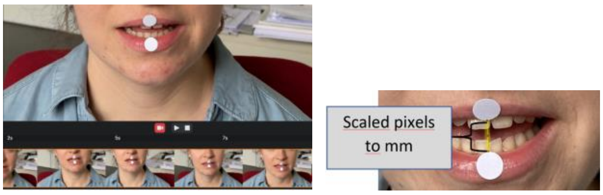

Figure 1. (Left) Video (top panel) of speaker producing a syllable /da/. In the bottom panel the video was parsed into individual motion capture frames at every 2 ms timestamp. (Right) Static image of motion capture frame where maximum lip aperture was calculated.

Table 3. Maximum lip aperture in mm for the individual vowels by speaker.

	a	e	i	o	u
Female	10.35	9.36	7.61	6.03	2.22
Male	7.77	6.98	12.08	12.22	3.65

Before constructing the models, all variables were standardized by removing the mean. All models were built using the same parameters: the initialization method is defined as "K-means++" (the first clusters are selected based on empirical probability distribution instead of using a random selection), the maximum number of iterations was set to 300, and the K-means algorithm used was Lloyd's algorithm [14]. Each model ran 10 times with different centroid seeds. To determine the most appropriate number of k, we used the elbow method. This is a common method to identify suitable cutoff values, since it suggests that values passed the elbow of the curve would not add more information. It consists of iterating over different k values and plotting the sum of squared distances at each cluster (Figures 2-3). In our experiments, we iterated the number of clusters from 2 to 10. The most suitable number of clusters is selected based on the elbow of the curve.

Furthermore, after identifying the most suitable number of clusters, we performed dimensionality reduction. This technique has been proven to increase the performance of a clustering algorithm [15] since it reduces data complexity. To do so, we employed Principal Component Analysis (PCA). This statistical method consists of transforming the data in an unsupervised manner into a smaller number of dimensions while conserving the data information. The simplified dimensions are referred to as components. Each component is the result of the linear combination of all variables. The first component consists of the normalized linear combination of those variables with the highest variance. Usually, the first components are able to contain all variance from the data. To identify the most suitable number of components in our dataset, we built a K-means model for the first component and for the first and second components.

Moreover, since the most suitable number of clusters might not be the same as the number of vowel categories, we repeteated the PCA analysis and trained a K-means model by setting k as the number of vowel categories (n=5).

Results of the models were evaluated based on adjusted rand index (ARI) and silhouette score. The silhouette score calculates how well each cluster is distinguished from the other clusters according to the ground truth. The score is useful to evaluate a clustering algorithm. It ranges from -1 to 1, where 1 refers to clusters being perfectly distinguished from each other and -1 means that all clusters were assigned wrongly. The ARI expresses the number of correctly assigned clusters. This metric requires ground truth values to measure the similarity of the different clusters with the true labels. It ranges from –1 to 1, for which the closest to 1 refers to a better clustered data. A score close to 0 indicates random performance. Silhouette score calculates how distinct clusters are from each other. A score of 1 corresponds to perfectly distinguished clusters.

3. MODEL OUTCOMES

3.1 Elbow method

Figure 2 shows that for the acoustic data the number of sum of squared distances decreases and creates a clear elbow in the cluster 6. For the visual and acoustic data (Figure 3), the elbow in the curve is shown until cluster 7. Therefore, for the acoustic data, models are built using 6 clusters and for the acoustic and visual data, models employ 7 clusters.

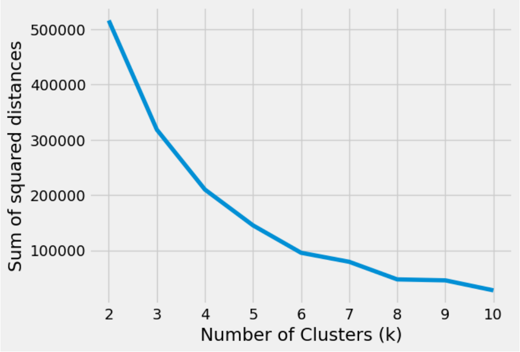

Figure 2. Sum of squared distances for each number of clusters for the acoustic data.

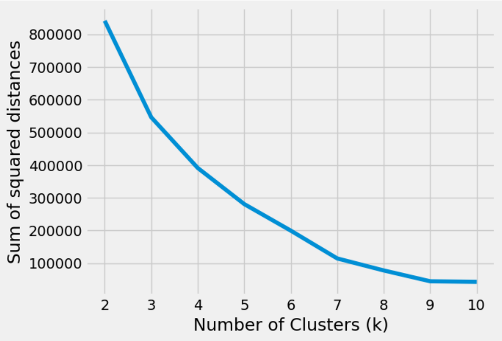

Figure 3. Sum of squared distances for each number of clusters for the acoustic and visual data.

3.2 Acoustic data

Table 4 shows the results obtained when building a model with two different components and two different clusters. As expected, the number of clusters identified from the elbow method shows the best-performance.
Results using 6 clusters and the first component, or the first two components are similar for the ARI score.
The highest silhouette scores are obtained when using the first component, which suggests that the first component is more suitable to clearly distinguish the different clusters. Figures 4 and 5 illustrate how the true labels are clustered using component 1 and component 2.

Table 4. Silhouette and ARI score using different number of components and number of clusters based on the elbow method and the number of ground truth labels.

Number of clusters	Number of components	Silhouette score	ARI
6	Comp. 1	0.64	0.34
	Comps. 1 and 2	**0.59**	**0.35**
5	Comp. 1	0.68	0.26
	Comps 1 and 2	0.58	0.25

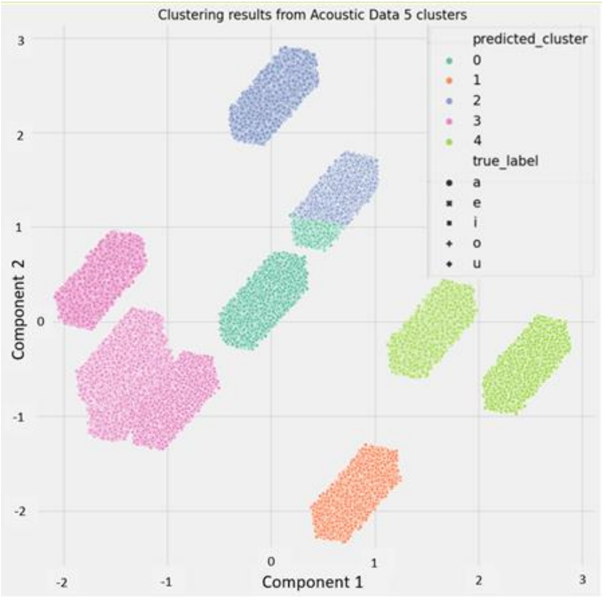

Figure 4. Figure showing the 5 predicted clusters against the true labels in component 1 and component 2.

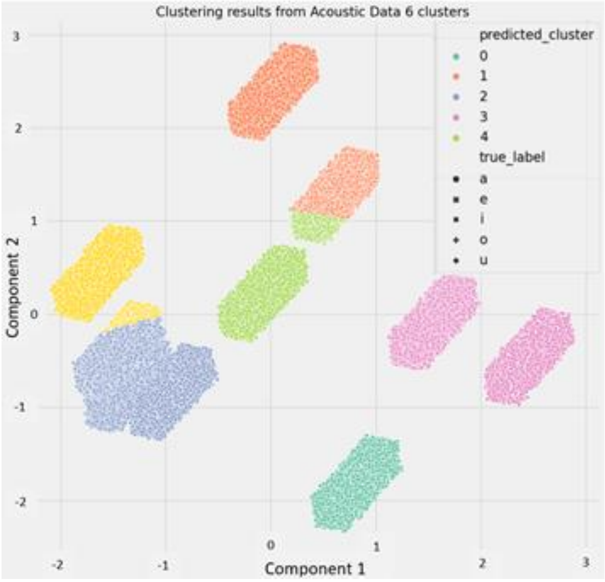

Figure 5. Figure showing the 6 predicted clusters against the true labels in component 1 and component 2.

3.3 Acoustic and visual data

Table 5 shows the results obtained when building a model with two different components and two different clusters for the acoustic and visual data. Using the first two components with 7 or 5 clusters shows the best ARI scores. The best distinguished clusters (silhouette score) are obtained using only the first component. Figures 6 and 7 illustrate how the true labels are clustered using component 1 and component 2.

Table 5. Silhouette and ARI score using different number of components and number of clusters based on the elbow method and the number of ground truth labels.

Number of clusters	Number of components	Silhouette score	ARI
7	Comp. 1	0.81	0.43
	Comps. 1 and 2	**0.73**	**0.45**
5	Comp. 1	0.82	0.38
	Comps. 1 and 2	0.64	0.45

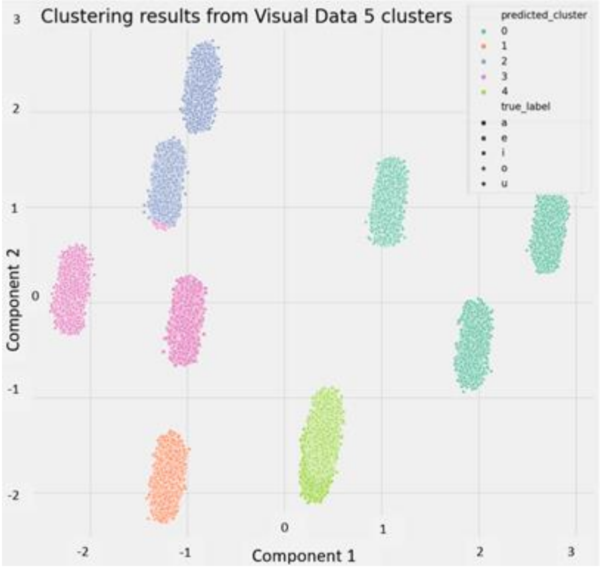

Figure 6. Figure showing the 5 predicted clusters against the true labels in component 1 and component 2.

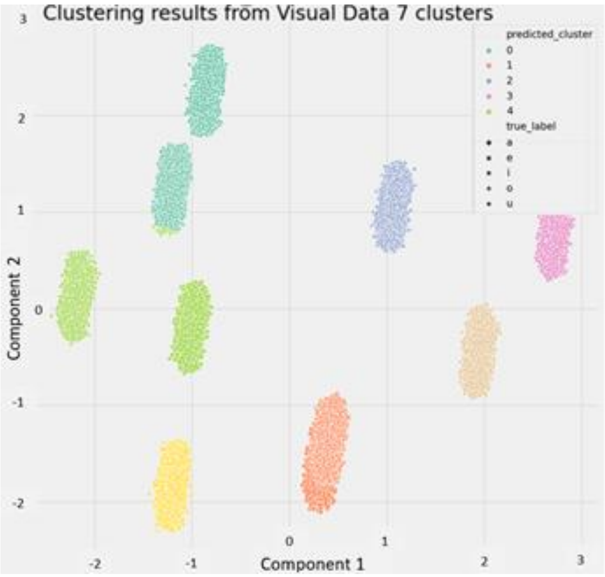

Figure 7. Figure showing the 7 predicted clusters against the true labels in component 1 and component 2.

Results indicate that using the acoustic data clusters the different vowel categories with a similarity (ARI) of up to 0.35 (Table 4). When examining the predicted clusters compared to the vowel categories (Figure 4), one can notice that [e] and [i] are clustered together but [i] is also clustered independently.

Moreover, it is important to note that [a], [o] and [u] are grouped together when only 5 clusters are used. However, when a model with 6 clusters is built, [a] is clustered independently while [o] and [u] are still aggregated.

When visual data are added to the acoustic data and 7 clusters are used, the confusion between [o] and [u] is resolved by identifying two separate clusters for [o] and [u] respectively (as illustrated in Figure 5). We surmise that one cluster refers to the male samples and the other to the female ones. However, when 5 clusters are used, the confusion between [o] and [u] remains as in the acoustic data. Regarding [a], [e], and [i], Figure 5 shows that these three categories are confused and mixed in three clusters.

Overall, our results indicate that adding visual data enhances the clustering algorithm, achieving an ARI of up to 0.45 when comparing the similarity of the predicted clusters to the ground truth labels (vowel categories).

4. DISCUSSION

We postulated, based on the results of our previous psychoacoustic perception tests, that the incorporation of visual information into an acoustic stimulus would aid in the recognition of vowel categories in the face of a noise source. The results in section 3 corroborate our underlying assumption that visual information, in tandem with an acoustic stimulus will increase response accuracy across subject groups. The results show that our models better classify vowels when a visual stimulus, in this case maximum lip aperture, accompanies the acoustic stimulus. This is an intuitive finding in that it suggests that humans may call on any informational channel at their disposal in order to fill in the gaps when a message is perturbed by noise. However, a few questions remain with regard to how the auditory and visual information interact, which we will deal with below.

The models we presented are basic prototypes that explain back vowel confusion across a wide range of subjects. Nevertheless, the models did not reflect two fundamental assumptions, which we will address in future studies. On one hand, the model did not take into account the fact that one of the groups (children with CI) receive a distorted input signal because of the device that would affect their learning and classification of categories. In future rounds this must be programmed into the model because the distorted input affects how a listener represents phonological classifications and their subsequent access. We propose here that this can be done with more sophisticated models, such as convolutional neural networks that can be trained on audio data, by computing the short-time Fourier transform. The way to simulate hearing impairment, thus, may be to reduce or alter the frequency spectrum of the spectrogram like CI programming.

Additionally, we did not take into account the fact that non-native speakers of Spanish, such as the EN-L1 group in [2] and [4] do not start learning from zero, as our model does. These subjects already have phonological categories from their first language programmed into their cognition. The real question here is how do speakers that already have first-language phonological categories codified in the mind map non-native stimuli to phonological categories in the face of different noise sources. This can be dealt with in future rounds of testing by the creation of a new model which is first trained on English vowels and later asked to classify non-native categories in noise.

Another question for future rounds of inquiry will be to ask how much information from the visual/auditory signals are being used by any one group to classify phonological categories. We assume that the relationship between the two information signals is dynamic (i.e. it is modulated as a function of time and context). That is, the information listeners extract from the auditory and visual information is not constant, and therefore not easily measurable.

An additional assumption is that the interaction between the visual and audio signals is modulated by the state of the original system, that is, the degree to which these informational signals interact is dependent on the system in which they operate (i.e., we would expect different levels of integration for children with CI than children with TH etc.). We plan to address these questions in future electrophysiological studies using electroencephalography data (EEG) and eyetracking.

5. ACKNOWLEDGEMENTS

This work was financed by a grant [ref. Pid2019-105929ga-i00] by the ministry of science and innovation (spain).

6. REFERENCES

[1] C.E. Shannon, “A mathematical theory of communication”, *The Bell System Technical Journal* 27, pp. 623–656, 1948.

[2] M. Gibson, M. Schlechtweg, B. Blecua Falgueras and J. Ayala Alcalde, Language-specific interactions of vowel discrimination in noise, in *Proc. of the International Speech Communication Asssociation (Interspeech),* (Incheon, South Korea), pp. 3118–3122, 2022.

[3] M. Gibson, M. Schlechtweg, J. Ayala, A. DiCaccio, X. Wang, and L. Xu, Energetic and informational masking effects on Spanish vowel discrimination, to appear in *Proc. of the International Congress of Phonetic Sciences,* (Prague, Czech Republic), pp. to be announced, 2023.

[4] H. Reetz and A. Jongmann, *Phonetics: Transcription, Production, Acoustics, and Perception*. Wiley-Blackwell, 2009.

[5] A.R. Bradlow, A comparative acoustic study of English and Spanish vowels. *Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1916–1924, 1995.

[6] T. Välimaa, T.K. Määttä, H.J. Löppönnen and M.J. Sorri, Phoneme recognition and confusions with multichannel cochlear implants: Vowels, *Journal of Speech, Language, and Hearing Research* vol. 45, pp. 1039–1054, 2002.

[7] M.E. Renwick, Vowels of Romanian: Historical, Phonological and Phonetic Studies, Phd dissertation, Cornell University, 2012.

[8] M.H. Mateus and E. de Andrade, *The Phonology of Portuguese*. Oxford, U.K: Oxford University Press, 2000.

[9] J.I Hualde, *Los sonidos del español*, Cambridge, U.K: Cambridge University Press, 2014.

[10] J. Rahilly, Vowel disorders in hearing impairment. in Ball, M. J., Gibbon, F. E. (Eds.), *Handbook of Vowels and Vowel Disorders,* Psychology Press, pp. 364–385, 2012.

[11] M. Molis and M Leek, Vowel identification by hearing-impaired listeners in response to variation in formant frequencies. *Journal of Speech, Language, and Hearing Research,* vol. 54, no. 4, pp. 1211–1223, 2011.

[12] P.J. Monahan and W.J. Idsardi, Auditory sensitivity to formant ratios: Toward an account of vowel normalisation, *Language and Cognitive Processes,* vol. 25, no. 6, pp. 808–839, 2010.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, A., D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[14] R. Ostrovsky, Y. Rabani, L. Schulman and C. Swamy, The Effectiveness of Lloyd-type Methods for the K-means Problem in *J. ACM* 59, 6, 2012.

[15] C. Ding and X. He, K-means clustering via principal component analysis. in *Proceedings of the twenty-first international conference on Machine learning (ICML '04)*. Association for Computing Machinery, New York, NY, USA, 29. https://doi.org/10.1145/1015330.1015408, 2004.

[16] E. Bonabeau, M. Dorigo and G. Theraulaz, *Swarm intelligence: from natural to artificial systems*. Oxford University Press, pp. 9–11, 1999.