EMOTIONAL VALENCE IS ENCODED BY THE PHONEMIC CONTENT – SOUND-SYMBOLIC EXPERIMENT

Velina Slavova, Ventsislav Dimitrov

Abstract: *This paper presents an experiment on emotional sound symbolism and reports the initial results. The study aims to determine whether listeners associate certain types of sequences of phonemes with positive or negative emotional tones. Participants are presented with visual and auditory stimuli to match them. Both types of stimuli are pre-categorized as positive or negative. The visual stimuli were realistic photographs. Auditory stimuli were nonsense sequences of phonemes presented within a game-like scenario, in which they are machine-pronounced without intonation and presented as being words in an unfamiliar language. These pseudo-words are phonemic sequences, generated as positive, negative, and neutral based on correlation coefficients between vowel-consonant pairs and emotional valence obtained in a previous study. Subjects matched the different types of sound sequences to photographs differently depending on the emotional tone embedded in the visual stimuli. This indicates that they perceive "emotional" differences in the sound sequences.*

Keywords*: Sound-symbolism, Speech emotion recognition and synthesis, Emotional valence*

ITHEA Keywords: I.2.7 Natural Language Processing, Cognitive simulation

DOI*: https://doi.org/10.54521/ijita30-03-p03*

Introduction

The study of emotions has been increasing recently, especially due to the need for machine recognition and machine speech generation. Emotions are known to impact human physiological, mental, and cognitive levels. Two basic models of emotions are used. The ***discrete*** model categorizes emotions with specific names (joy, sadness, anger, etc.), with variations in the number and composition of emotions highly dependent on the researchers. The second, ***continuous*** model, is called VAD (Valence-Arousal-Dominance), and represents emotional states on a spatial scale, most often along two axes: Valence, indicating positivity or negativity, and Arousal, indicating the level of agitation associated with the emotion. What is described here is based on the second model because of its universality and the abundance of evidence supporting its adequacy.

Recent research in the field of "sound symbolism" has explored the relationship between phonetic composition and the meaning of words. Relatively few, however, have found links between phonemic composition and emotion.

A study by Kawahara and Shinohara (2012) revealed that vowel-consonant pairs presented as acoustic stimuli evoked both a sense of magnitude and emotion in subjects, suggesting that sublexical components of speech convey emotional content through their phonetic characteristics '[Kawahara & Shinohara (2012)]. This discovery led to the hypothesis that sublexical components of speech, which are parts of words, convey emotional content through their phonetic properties. This laid to apply a Gestalt approach to the text, which views the text as a unified whole consisting of interconnected levels — sentences, words, and word parts — each contributing to encoding

emotional valence [Slavova, 2021]. The application of the Gestalt approach to text in recent years has yielded significant insights into the relationship between text and emotion. For instance, a series of corpus studies conducted on the German language has demonstrated that all levels of language, including syllables, convey emotional content (e.g., [Aryani et al., 2016]). This entire reasoning underpins the Gestalt approach to phonemic content used here, viewing the speech as a cohesive entity comprising interconnected levels — sentences, words, and word parts — each playing a role in encoding emotional valence.

Used preliminary results

Exploring the relationship between the sublexical level of a text and its emotional valence has yielded promising results, particularly in the correlation between specific vowel-consonant pairs and text valence [Slavova, 2019]. A corpus analysis [Slavova, 2020, 2021] of the EmoBank dataset [Buechel & Hahn, 2022], containing 10,000 valence-evaluated English sentences across seven genres, further investigated this relationship. Phonetic transcriptions of words, following British standards, were analyzed at the sublexical level down to consonant-vowel pairs, known as ***biphones***. These combinations were chosen because they are fundamental to word formation in speech.

The results confirmed that emotional valence is indeed encoded at the sublexical level. Principal component analysis revealed that the second principal component of biphone frequency data correlated strongly ($r = 0.96$) with readers' valence ratings [Slavova, 2021]. The analysis also showed

significant differences in biphone frequencies between positive and negative sentences, and these correlations were used in further studies.

The validity of these correlations in texts outside the corpus was also tested. Pearson correlation coefficients between biphones and valence, calculated from EmoBank data, were applied to randomly gathered news items from the Internet. The result further supported the hypothesis that emotional valence is encoded in phonemic composition [Slavova & Andonov, 2022].

However, these statistical results do not explain the underlying cause of this phenomenon. How did valence-related biphones enter the text independently of its meaning?

One important question remains: do the biphones themselves evoke specific emotional responses? The experiment presented here aims to determine whether the mere hearing of a series of biphones of positive or negative type evokes a feeling of a particular emotional valence of the perceived sound message.

General presentation

The purpose of the experiment described here is to test whether, purely perceptually, hearing a series of particular biphones is associated with positive or negative emotional valence. The approach taken was to have the subjects match a visual and an auditory stimulus. The general scheme of the experiment is illustrated in Figure 1.

Realistic photographs serve as visual stimuli. They are in two main categories - positive and negative. The pictures are fed to the subjects on a screen.

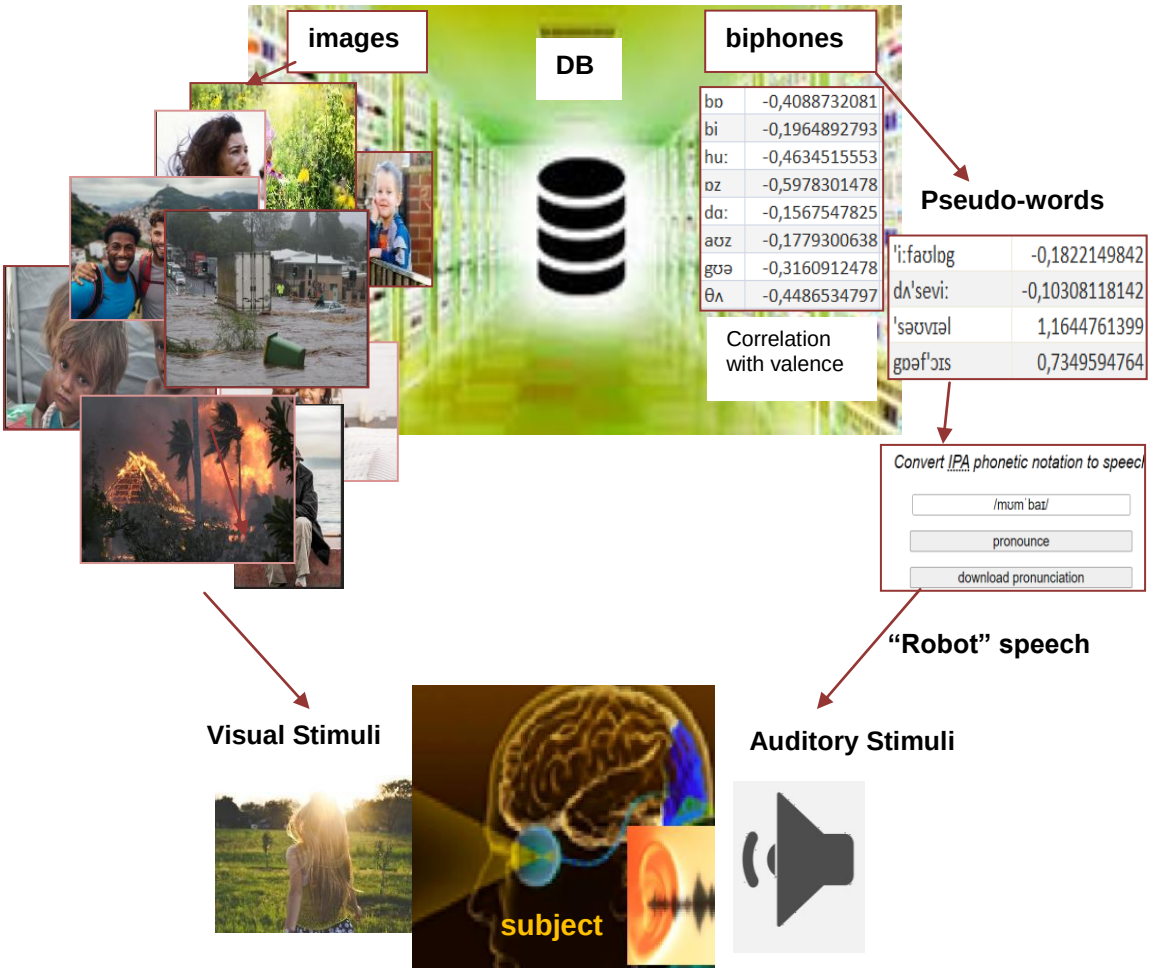

Figure 1. General scheme of the experiment

Auditory stimuli are pseudo-words obtained by generating random sequences of biphones. For the specific purpose of the study, the utterance of the sequences of biphones must be devoid of intonation and prosodic elements marking

emotional speech such as pauses and elongations. The pseudo-words are therefore submitted to "*phoneme synthesis*" [phoneme synthesis], an available online resource that parses IPA phonetic notation and converts it to speech. The result is a machine-pronounced sequence of phonemes devoid of intonation, likened here to "robot speech." It is presented to the subjects as a sound stimulus.

The task of the participants is to match visual and auditory stimuli. Stimuli are presented randomly according to a specially designed game-like scenario. Subjects are assumed to distinguish and properly react to the emotional valence inspired by the visual stimuli.

Materials and procedures

Visual stimuli

A carefully curated set of images was selected by the experiment developers, consisting of 25 images that convey negative emotions and 25 images that inspire positive emotions. Additionally, each image underwent a thorough manual assessment by 10 independent evaluators. They scored the images on a scale of 0, 0.5, and 1, based on how accurately they believed the image communicated the intended positive or negative emotional valence. Only images with an average score exceeding 0.8 were included in the experiment, resulting in a final set of 12 negative and 14 positive photos.

These selected images were further categorized based on their primary content into two types: images predominantly featuring human faces and images depicting landscapes. This resulted in 15 images of human faces (with 8 positive and 7 negative) and 11 images of landscapes (with 6 positive and 5 negative).

Auditory stimuli

The list of the nearly 800 biphones, found in EmoBank, was categorized based on their "weights" which are the correlation coefficients (Pearson) with valence, derived in the analysis of the sublexical level of the corpus [Slavova, 2020]. The weights of the biphones are in the range -0.8, 0.8. For the random generation of pseudo-words, the biphones are separated into three distinct groups: negative (weight less than -0.15), neutral (weight ranging between -0.15 and 0.15), and positive (weight greater than 0.15). Leveraging this sorted list, a JavaScript program was created to randomly generate a total of 630 pseudo-words. The program additionally assigned a randomly chosen stressed biphone. For each pseudo-word, a "weight" was calculated as the sum of the correlation coefficients of the constituent biphones, as illustrated in Figure 1.

The randomly generated sequences formed 9 groups of pseudo-words as follows:

This approach ensured a diverse and balanced array of pseudo-words, evenly distributed across the various biphones' combinations for robust analysis.

Not all phoneme sequences are suitable for human pronunciation. To enhance the similarity with human speech, the machine-pronounced pseudo-words were evaluated by two language advisors, who manually selected 9 pseudo-words

from each group that looked more like pronounceable words in a human language. Thus, the collection of pseudo-words used in the experiment contained 81 random biphonic sequences.

Table 1. Groups of pseudo-words and number of generated sequences

	3 biphones	4 biphones	5 biphones
	(e.g. ***gʊə'θʌɒt***)	(e.g. ***ɔːj'teəɑːʒlɔɪ***)	(e.g. ***uːləknə'ɔːwɔːg***)
negative	70	70	70
neutral	70	70	70
positive	70	70	70

The pseudo-words containing 3 and 5 biphones were used directly as audio stimuli in the experiment. For pseudo-words containing 4 biphones, 6 "sentences" were manually constructed — 3 with positive pseudo-words and 3 with negative pseudo-words.

All data, including images, pseudo-words, and "sentences", were organized and imported into a database, establishing a comprehensive foundation for the subsequent phases of the experiment.

Scenario of the experiment

The experiment scenario was designed with user-friendliness in mind, ensuring that it is not only time-efficient but also enjoyable for the participants. The experiment is anonymous. The procedure begins by collecting the native language and gender of each participant to allow a corresponding analysis of the results.

To frame the experiment in an engaging, game-like manner, participants are asked to imagine a scenario where their friend speaks an unfamiliar language, necessitating the use of a translation device, but the device broke down. The translator’s failure, it is explained, consists in repeating the words as they are in the original language, without any intonation or emotional inflection. To acclimate participants to the perception of such speech, they are first provided with an example of the translator's output, setting the stage for the subsequent sections.

In the first part of the experiment, which involves matching sounds to images, participants are asked to imagine they are flipping through a photo album with their friend. As the friend shows them each image, he makes a short comment in his language, but the comment is delivered by the robotic voice of the damaged translator. The participant's task is to match the correct word — chosen from a set of three options — to the image, as shown in Figure 2.

On four successive screens similar to the one shown in Figure 2, participants are shown a randomly selected image, alternating between positive and negative content, including both faces and landscapes. Alongside each image, three random pseudo-words are presented: one negative, one neutral, and one

positive. Participants can listen to the pseudo-words as many times as they like before selecting the one they believe best matches the displayed in the image.

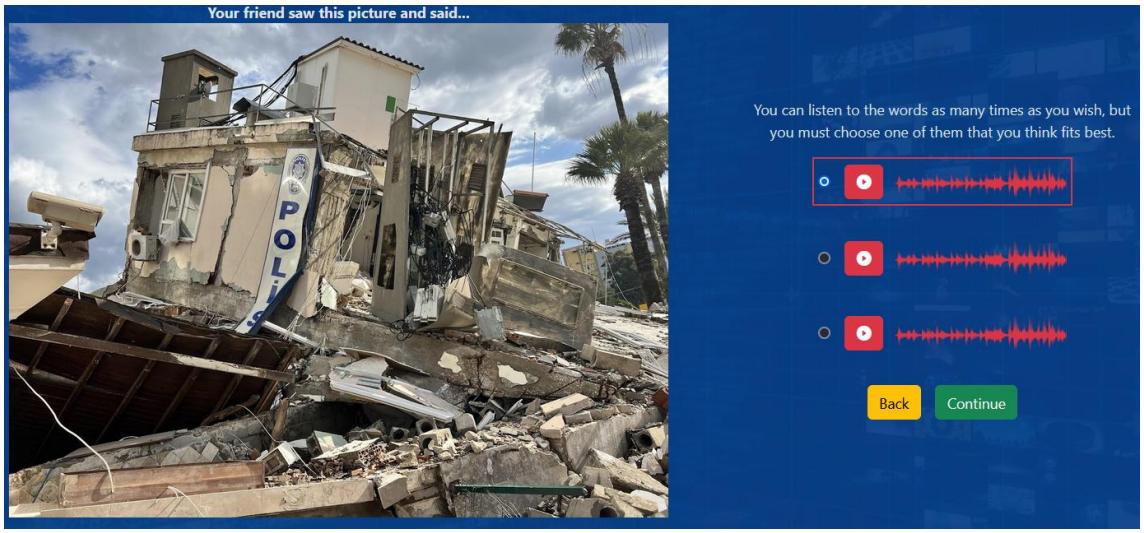

Figure 2. A screen from the sound-to-image matching part of the experiment

In the second section, which involves image-to-sound matching and serves as a control, the task is reversed. As shown in Figure 3, participants are presented with a positive or negative 'sentence' made up of pseudo-words and must choose between a positive and a negative image. The scenario remains within the same game-like environment, where the friend says something, and the participant needs to select the corresponding image. Sentences are randomly sourced from the database. On the first screen, participants choose between a positive and a negative image featuring faces, while on the second screen, they choose between positive and negative landscape images.

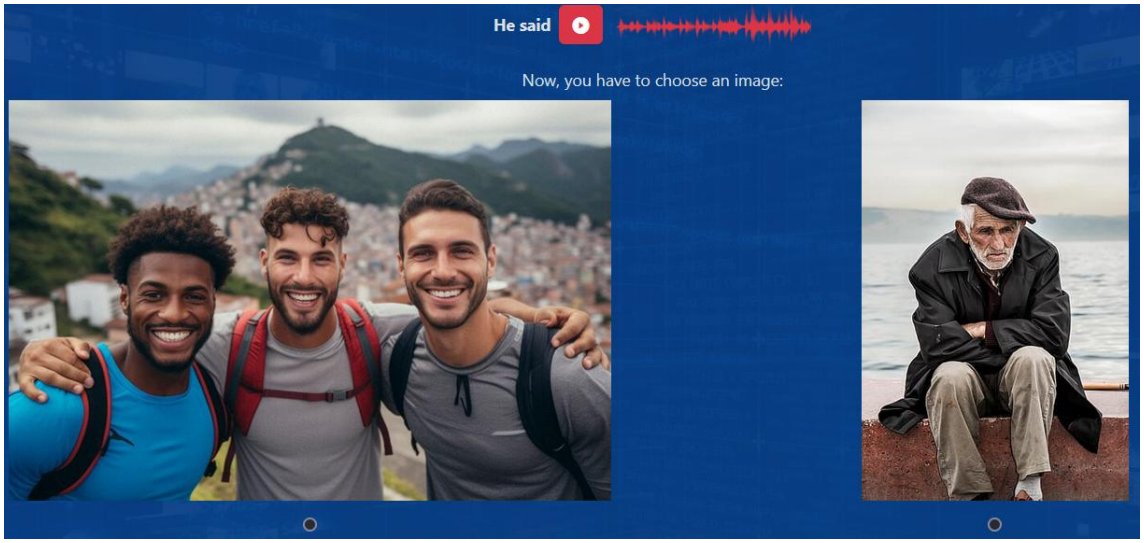

Figure 3. A screen from the image-to-sound matching part of the experiment

On the final screen, participants are shown their scores. They earned 2 points for correctly matching an image and a biphonic sequence, 1 point when having selected the neutral pseudo-word, and 0 points for a mismatched correspondence. This scoring system provides immediate feedback and fosters a sense of achievement, motivating participants to retry for a higher score, which helps gather more statistical data.

This thoughtfully crafted scenario is designed to be quick, easy to follow, and enjoyable. The final scoring system incentivizes participants to engage more deeply, thereby contributing to a richer dataset.

Technical aspects of the realization

With the dataset and experimental scenario prepared, the team initiated the development of the web application intended for participant engagement in the study.

For the backend architecture, a REST API was developed utilizing PHP in conjunction with the Symfony framework and a MySQL database to manage the dataset. The initial dataset was organized into four distinct entities: an Image entity representing each previously selected image, a Biph entity containing all unique biphones in the experiment, a Word entity derived from the Biph entity for each previously selected pseudo-word, and a Sentence entity constructed from the Word entity, encompassing the chosen "sentences." In addition to these foundational entities, three more entities were prepared to record participant data. The experiment's two distinct components were separated into two entities: a Screen entity and an ExtraScreen entity. The Screen entity comprises an Image and three Words for the sound-to-image matching task, whereas the ExtraScreen entity includes two Images and one Sentence for the image-to-sound matching task. Finally, a Survey entity was constructed to encapsulate data for each experiment attempt, incorporating four Screens, and two ExtraScreens, as well as metadata such as the start and end times of the attempt, the score achieved, and demographic information including the participant's language and gender. Given that no data uniquely identifies the participants, there is no necessity for a distinct entity for user profiles. This recursive entity construction methodology ensures minimal data redundancy, helping the subsequent data analysis.

The API exposes three endpoints: one for initializing a new Survey, one for concluding a Survey, and an administrative endpoint for exporting the aggregated data. To initialize a new Survey, the endpoint requires only the participant's language and gender. Based on the predefined experimental scenario, new unique Screens and ExtraScreens are dynamically generated by randomly selecting Images, Words, and Sentences.

A frontend interface was developed using Vue.js to present the data and communicate with the backend API. The design chosen was minimalistic, prioritizing visual and auditory stimuli while maintaining an intuitive and accessible user interface. To achieve the synthesis of pseudo-word pronunciations, an external library was included, thereby completing the application. Preliminary testing indicated that the system operated as expected.

The running system can be accessed at the following URL: https://cscb634.pwpwebhosting.com/

Upon conducting the experiment's initial trials, the development team was uncertain about the results. The robotic articulation of all pseudo-words appeared quite similar — extraordinarily strange and fairly unpleasant. This raised concerns about whether the experiment participants would make any difference when selecting a corresponding item or would perform the tasks by "voting" randomly.

First results

As of the publication of this report, 120 trials have been conducted on the experiment site, with 85 completed and included in the statistical analysis. The participants comprise 40 women and 45 men, representing 15 different native languages. Since this is a preliminary analysis, we are presenting descriptive statistics.

We will begin by discussing how participants perceive the synthetic robot speech. As noted, while the machine-generated utterances were accurate in composition, the participants often described the robot-talking as unpleasant

due to its distinctly artificial 'machine voice' without intonation. In the second part of the experiment, depicted in Figure 3, participants engaged in image-to-sound matching as a control task. In this phase, participants listened to a sequence of 12 biphones that mimicked words in a sentence and had to choose between two images — one positive and one negative. This setup gave them a 50% chance of selecting the 'correct' image. The results from the 85 completed trials are illustrated in Figure 4.

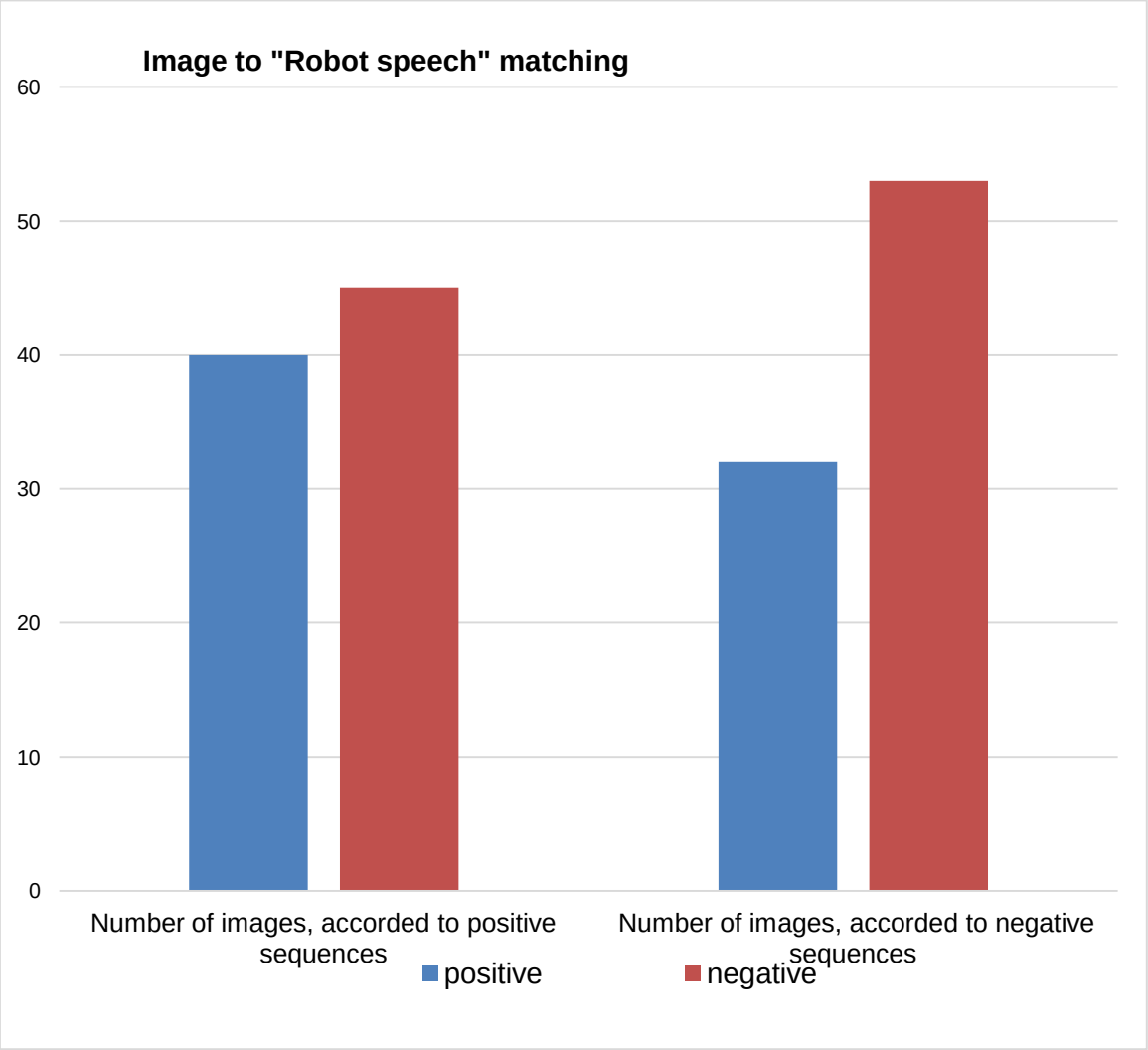

Figure 4. Number of positive and negative images matched to positive and negative ‘sentences’

As shown, participants generally perceived long sequences of both positive and negative biphones spoken with the machine-uttered speech as unkind. Remarkably, though, negative sequences elicited a significantly stronger negative response compared to positive sequences.

In the first part of the experiment (Figure 2), participants engaged in sound-to-image matching by selecting one of three pseudo-words to correspond with the visual stimulus. This setup provided each participant with a 33% chance of choosing any given pseudo-word. The result based on the 85 full tests performed is presented in Figure 5.

As presented in the plot, the selection of the corresponding pseudo-words in response to positive and negative visual stimuli shows clear distinctions. Negative visual stimuli evoke an emotional response that aligns more with negative pseudo-words. It is not the same for positive visual stimuli. Interestingly, for positive stimuli, participants frequently (almost 50% of the time) selected pseudo-words with low emotional weight (neutral). This pattern warrants further investigation and experimentation; at this stage, it remains speculative.

The primary goal was to determine whether emotional valence can be perceived in biphone sequences. Even though all pseudo-words sound similar — being unfamiliar, surprising, lacking intonation and pauses, and incoherent with any known language — it appears that participants involuntarily detect some form of difference. This distinction is likely due to the biphonic composition of the pseudo-words. Therefore, the answer to the main question is clear: yes, emotional valence is indeed perceived by people in biphone sequences.

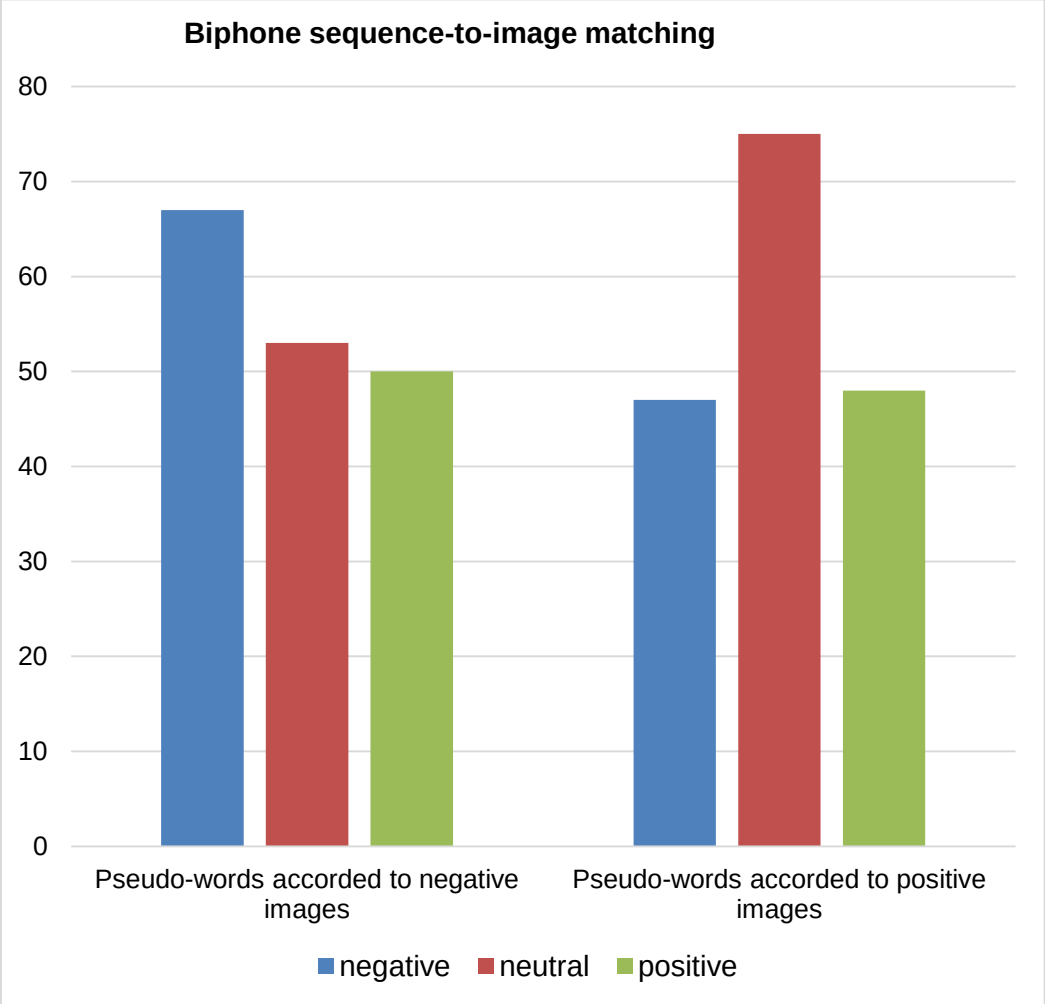

Figure 5. Number of pseudo-words of different types matched to positive and negative images

Conclusion

The experiment results indicate that people detect emotional differences in the phoneme sequences.

The phonemic content of words in a language, therefore, is not arbitrary, as the correlation coefficients of the biphones are derived from the words in authentic

sentences. This suggests that words have evolved in the language development process to reflect, through sequences of phonemes, both the meaning AND the emotional connotations of the concepts they signify.

It should be noted that the biphonic correlations with valence were derived from English texts. Participants in our experiment were not native English speakers. This provides a basis to hypothesize that the observed perceptual phenomenon may be universal.

The sample is balanced in terms of gender. Still, the majority of participants (63) are native Bulgarian speakers, and many other languages are represented by only one participant each. This makes it impossible to conduct a meaningful analysis based on native language at this stage.

Further data analysis and additional experiments are needed.

Discussion

There is an increasing necessity for systems that can recognize emotional tones and can generate speech messages with a specific character. While the relationship between prosodic features of speech and emotional tone is relatively well-studied, little is known about how linguistic messages' phonemic composition affects the perception of emotional tone. The same meaning of a message can be conveyed using different words, that is – different phonemic sequences. Thus, it can inspire different emotional dispositions. The result of the experiment presented here, along with its preliminary results, indicates that people's emotional state is perceptually influenced by phonemic composition.

This fact should be taken into account, and further, more precise studies are needed.

ACKNOWLEDGMENTS

This system was developed within the framework of the programming and database practice included in the BSc program in Informatics at the New Bulgarian University. All subtasks arising in the development of the system were developed by third-year undergraduate students in Informatics. The authors express their gratitude to Hristina Zareva, Darina Aleksieva, Yana Ivanova, Antoan Paskalev, and Anton Staykov.

Bibliography

[Aryani et al, 2016] Aryani, A., Kraxenberger, M., Ullrich, S., Jacobs, A. M., & Conrad, M. Measuring the basic affective tone of poems via phonological saliency and iconicity. Psychology of Aesthetics, Creativity, and the Arts, 10(2), 191. https://doi.org/10.1037/aca0000033

[Buechel and Hahn, 2022] Buechel, S., & Hahn, U. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. arXiv preprint arXiv:2205.01996.

[Kawahara & Shinohara, 2012] Kawahara, S., & Shinohara, K. A tripartite trans-modal relationship among sounds, shapes and emotions: A case of abrupt modulation. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 34, No. 34). https://escholarship.org/uc/item/47b452vw

[Slavova & Andonov, 2022] Slavova, V., & Andonov, F. Bad news or good news when recognizing emotional valence using phonemic content. In 2022 21st

International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE. https://doi.org/10.1109/INFOTEH53737.2022.9751339

[Slavova, 2019] Slavova, V. Towards emotion recognition in texts–a sound-symbolic experiment. International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE), 7(2), 41-51 https://doi.org/10.5937/IJCRSEE1902041S

[Slavova, 2020] Slavova, V. Emotional valence coded in the phonemic content–Statistical evidence based on corpus analysis. Cybernetics and Information Technologies, 20(2), 3-21. https://doi.org/10.2478/cait-2020-0012

[Slavova, 2021] Slavova, V. On the revealing the emotional valence in communication by text. Procedia Computer Science, 192, 1514-1523. https://doi.org/10.1016/j.procs.2021.08.155

Internet Resources

[phoneme synthesis] https://itinerarium.github.io/phoneme-synthesis/

Authors' Information

Velina Slavova *– prof. in computer science at New Bulgarian University; e-mail: vslavova@nbu.bg*

Major Fields of Scientific Research: models and cognitive approach to AI

Ventsislav Dimitrov *– student in computer science at New Bulgarian University; e-mail: f112739@students.nbu.bg,*

Major Fields of Scientific Research: computer science