Prose2Sound: Sonification of Word Embeddings

Calvin Lu cl3644 & William Gu xg2355

***Abstract*—Visualization of data is typically limited to two or three spatial dimensions, but with sonification, it may be possible to represent more information by encoding data into sound. In this work, we look into finding mappings from word embeddings to sound such that the generated sounds are audibly differentiable and retain semantic features about the encoded words. Apart from artistic value, sonification may have use as an alternative method of representing data; patterns that aren't discernible visually might be discernible aurally.**

I. Introduction

Prose and lyrics contain an incredible amount of semantic information and complexity, but we lack a way to break down or analyze these complex features easily. Music and sound, on the other hand, along with their temporal characteristic, have proven to be uniquely useful in the exposure of certain patterns and analysis. Many recent advances such as text-to-image synthesis are the product of cross-domain pollination, so perhaps it is possible to leverage the sonification of word embeddings to reveal previously unknown patterns in the data. Our hypothesis is that by translating information between these mediums, we may be able to "see" language in a different light.

t-SNE and Principal Component Analysis (PCA) are two popular ways of representing data in lower dimensional spaces. These visualizations are used to help researchers notice broader trends in data; however, such methods of visualization are usually limited to 2-3 dimensions, which in turn limits the amount of information that can be expressed. We looking at the explained variances of the principal components in the PCA of word embeddings from language models such as W2V and BERT, we observe that the first 2-3 components explain less than a quarter of the variance in the data. This is expected, as these language models are already condensing extremely high-cardinality one-hot encoded tokens into a mere several hundred dimensions. Sound not only frees us from the dimensional limitations of cartesian space, it gives us access to the additional feature of time. It offers the potential for representation of data in unique ways, as it is possible to control not only the pitch, loudness, and timbre of the sound, but also mimic the perception of spatial dimensions via stereo sound.

Our new method, Prose2Sound, is one such way to represent word embeddings. Our work focuses on poetry and prose, but the method is applicable to standard language. By encoding the magnitudes of principal components into different pitches that are scaled by explained variance, we can represent significantly more information about each embedding than is possible with traditional visualization techniques.

II. Related Work

There is very little actual research in this domain. In past proceedings of IJCAI, we found only two [1] [2] weakly related articles in the past decade. In the sound domain, research is mostly focused on separation, event and anomaly detection, as well as speech synthesis. We also found very little work in the visualizations and explainability of language models that goes beyond 2D or 3D representations. We checked the past proceedings of ICASSP, Triple AI, ICML, and ICCC all with similar results.

This begs the question of why this particular task has not been investigated. Our best guess is that the practical applications of this task are not immediately clear and there is no easy answer for how to construct the map of word embeddings to sound, and how to compare these word embeddings and put them in conversation with each other. Nonetheless, we feel that the problem is interesting enough to investigate as the subject of our final project. Thus, we proceed given the understanding that there is little to no pre-existing research for the task of direct sonification of word embeddings. With that said, our work is built upon and inspired by prior work on language models, dimensionality reduction, and the applications of visualization in inspiring developments in deep learning research.

- **Language Models:** Word2vec [3] and BERT [4] are two popular language models that can be used to generate word embeddings. Word2vec produces context-independent embeddings: all senses of a word have the same representation. Because Word2vec ignores polysemy, we look to BERT for representations that can capture different meanings of the same word in different contexts.
- **Dimensionality Reduction:** t-SNE [5] and PCA are two popular methods of representing data in lower-dimensional spaces. While t-SNE is extremely popular for visualization, its emphasis of mapping neighboring points together may be less useful in the word embedding space as the magnitude of a vector may be heavily influenced by its frequency in the model's training corpus. On the other hand, PCA offers a deterministic solution that allows us to explain the most variance given a restricted number of components.
- **Applications of Visualization:** t-SNE visualizations of hidden layers in a CNN led to the discovery that adversarial perturbations also shift a perturbed image's representations in the hidden layers of the network [6]. Motivated in part by qualitative observations of these visualizations, the researches developed a defense against adversarial

examples by introducing an additional loss term to draw the hidden representations of perturbed images closer to those of their true class. With sonification, it may be possible to construct alternative higher-dimensional representations of information, which could potentially pave the way for future research into how hidden layer activations change with respect to changes in input or model architecture.

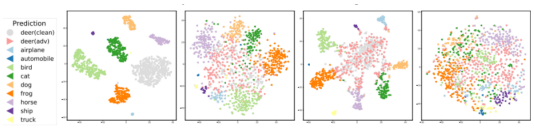

Fig. 1. t-SNE visualization of hidden representations of undefended model in comparison to three defenses

III. Methodology

General Approach: In Prose2Sound, text is first fed through a language model such as Word2vec or BERT to obtain language embeddings. As the embedding space spans hundreds of dimensions, direct sonification of embeddings would involve hundreds of pitches, which will likely make differentiation between words quite difficult. To get around this, we use PCA to transform the embeddings and attempt to utilize as many components as possible while maintaining audible differentiability between words.

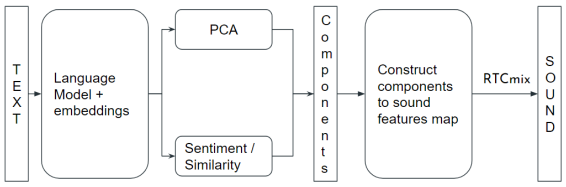

Fig. 2. Prose2Sound Architecture

The resulting vector, composed of the top k principal components of the original embedding, can then be mapped to sound. Each principal component is mapped to a pitch, and the amplitude of each pitch is scaled by the explained variance of the corresponding component. This method centers attention on the most important components; removal of this scaling factor results in excessive attention being diverted to random changes in the magnitudes of less important components. Additional features can also be added to make different words sound more distinct, such as by scaling sound duration by word length or adding additional dimensions to correspond to sentiment.

IV. Implementation and Technical Challenges

A. Word Embeddings

In this project, we made use of two Word2vec models: one pretrained on the Google News Dataset, and the other trained on the Gutenberg Poetry Corpus. Due to the relatively unusual expression of language in the poetry corpus, the embeddings from the specially-trained model were much better suited for our task. Because Word2vec does not discern between different senses of a word, we also look into BERT embeddings, which we define as the activations of the second-to-last layer. Although pretrained BERT assumes a sentence structure that may be excessively rigid for poetry and has a much larger embedding size, the amount of variance explained by the first handful of principal components is much higher than that of pretrained Word2vec. While BERT should be able to represent more semantic information, we may also need more dimensions to capture the same amount of explained variance. It may be too difficult for the human ear to distinguish between a large number of pitches, so we will largely stick with lower-dimensional Word2vec representations in our experiments.

B. Dimensionality Reduction

We elected to use PCA for dimensionality reduction due to its deterministic linear mappings which minimize the loss in explained variance from excluding components. We measure the amount of variance explained as a function of number of components to evaluate the information lost by representing data in lower dimensional spaces.

C. Sound Generation

The resulting vector of components is then used to determine the parameters for sound synthesis, such as frequency, amplitude, duration, and timbre. The normalized principal components are mapped to pitch, and the amplitude(volume) of each pitch is scaled by the explained variance of each corresponding principal component. For ease of differentiation, the duration of the each word's resultant sound is modified to correspond to an estimate of its syllable count. These parameters are then translated into a "scorefile", which is used by RTCmix to perform the actual synthesis of sound.

V. Experiments

Dimensionality Reduction: We conduct experiments to measure the proportion of explained variance in the top k principal components, as this gives us a method of evaluating an upper bound for the amount of information that can be expressed using our method. PCA is fit using the vocabulary from the Gutenberg Poetry Dataset, which skews the explained variance in the top k components in favor of models trained or fine-tuned on said dataset.

Sound Generation: Although our experiments in sound generation cannot be played in this report, generated output can be found in the project repository. These experiments include sonification of poetry and prose, pairs of similar words, and sentences composed of repeated and random words.

A. Results

The sum of explained variance in the top k principal components gives us an upper bound for the amount of information we can express using only those components. As

the components are fit using the vocabulary from the poetry corpus, it is no surprise that the Word2vec model trained on said corpus is able to capture the same variance in fewer components.

TABLE I
RETENTION OF EXPLAINED VARIANCE

Model/Dimensionality	10	20	30	40
W2V(Pretrained)	17.35%	26.13%	32.74%	38.19%
W2V(Gutenberg Poetry)	41.55%	57.74%	66.54%	72.86%
BERT(Pretrained)	30.80%	38.90%	44.03%	47.98%

Pretrained BERT captures more variance in fewer components relative to pretrained Word2vec, which is surprising as BERT's embedding representations contain 768 dimensions to the Word2vec model's 300 and encode information about context to differentiate between different word senses.

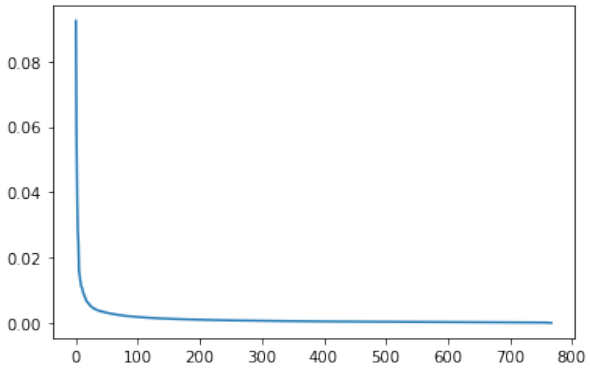

Fig. 3. Falloff of explained variance in BERT principal components

As we can see, the eigenvalues of the principal components do not drop off until tens of components in; visualization in 3 or less dimensions results in significant information loss for this application.

Ironically, we use visualization aids–spectographs–to compare the resulting sounds from two semantically similar words with embeddings generated from the Word2vec model trained on poetry. At each timestep, we increment the number of components included, so we see an increase in pitches with time. Perhaps as the result of training using a poetry corpus, "odor" and "smell" sound noticeably different–odor might have a much more negative connotation in this corpus.

In sonification of Ars Poetica by Pablo Neruda, we see some repetition as certain words are frequently repeated; however, it may take some familiarity with the specific language model and sound generation method to find nontrivial patterns in the data. Although not shown, qualitative analysis of BERT embeddings in the github repository reveals potential for our method: a sentence composed of nothing but the word "smurf" repeats roughly the same three sounds in a cyclical manner. Standard visualization might not capture this trend due to the inherently high dimensionality of word embeddings, and quantitatively finding the same pattern would involve

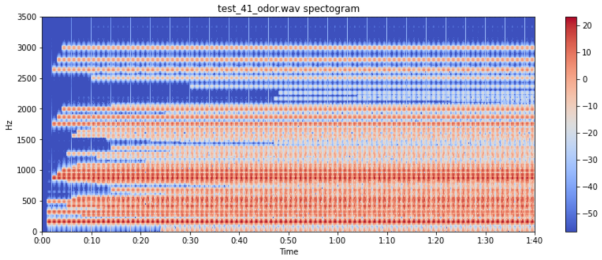

Fig. 4. Spectograph of "odor"

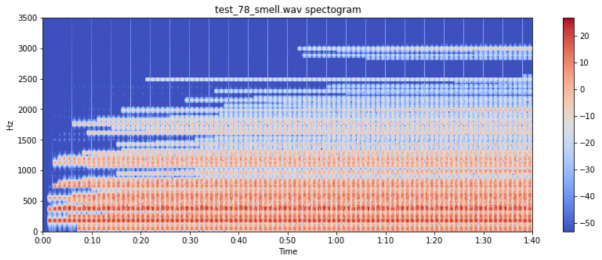

Fig. 5. Spectograph of "smell"

Fig. 6. Comparison of semantically similar words

painstakingly calculating and comparing similarities between every pair of words in each sentence.

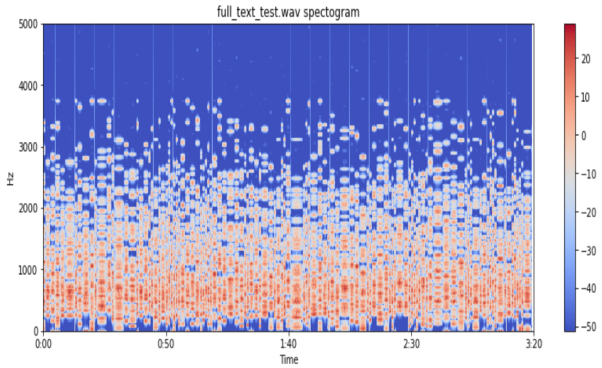

Fig. 7. Spectograph of poetry (Ars Poetica)

VI. CONCLUSION

Quantitative evaluation of the success of our method is difficult as it is an abstract task to concretely measure how well the words have been translated into sound. We also qualitatively discuss and evaluate generate sounds, even though this methodology may be fundamentally subjective andhighly dependent on musical exposure.

While the outputs of our method sound far from pleasant, it does appear that the relative distances between words are preserved as different words sound fairly distinct. Unfortunately, without extensive familiarity with both the structure of embeddings and sound generation technique, it may be difficult to identify meaningful patterns in the audio. While this method allows for higher-dimensionality representations

of data, higher dimensions can become incredibly noisy, and scaling the volume by additional components by explained variance may make it difficult to differentiate between more subtle differences.

Future work may seek to use higher quality embeddings that are more suited to the task; our approach starts with poetry and prose, but it may be easier to build upon standard language expression. There is also the possibility that other types of language models are more suitable to being sonified. Or it may also be beneficial to explicitly encode sentiment or part of speech.

Another major area of future development should be a smarter way of learning the relationship between language and sound. Instead of manually constructing the mapping between the two, as we have done, it is worth considering the possibility of a more elegant embedding technique, such as one that can learn a correspondence function between the two. Embeddings that include both sound and word information side by side, can be used to train such a model. See the use of deep symmetric structured joint embeddings (DS-SJE) from the text-to-image generation domain. [7]

REFERENCES

[1] Rafal Rzepka and Kenji Araki. Haiku generator that reads blogs and illustrates them with sounds and images. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 2496–2502. AAAI Press, 2015.

[2] AnneMarie Maes. The scaffolded sound beehive. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 2480–2481. AAAI Press, 2015.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[6] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness, 2019.

[7] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning deep representations of fine-grained visual descriptions, 2016.