Entropy shows: is it real speech, or a clone?

Terese Anderson[1], Grandon Goertz[2], Evan Ashworth[2]
[1]University of Chicago, US
[2]University of New Mexico, US

https://doi.org/10.36505/ExLing-2024/15/0003/000628

Abstract

This is an exploratory study that compares the conditional entropy formant values of naturally spoken words to the conditional entropy formant values of cloned vowels. It was hypothesized that cloned vowel formants, F1, F2, and F3, would have measurable and distinct differences compared to the formants of natural speech. This study shows that there are indeed variations in the entropy of Artificial Intelligence (AI) cloned vowels. These differences would be useful for forensic analyses and for distinguishing natural speech from AI generated imitations.

Keywords: entropy, cloned speech, forensics, formants, AI

Introduction

Voice cloning is a product of AI technology that creates a digital copy of a person's voice. In some cases, clones have been noted as sounding very close to or even identical to naturally spoken speech. This research compares naturally spoken English words to their cloned copies to answer the question: Do the entropy values of the formants of cloned vowels show measurable, identifiable differences compared to entropy values of the formants of natural speech?

Entropy concepts and calculations are used extensively in Information Theory and machine learning, and entropy computational methods can be used in evaluating phonetic phenomena. Entropy calculations are used to detect patterns in data that are related, and to show data relationships by providing numerical values that show the relative degree of overlapping information (see Haglund, Jeppsson, & Strömdahl, 2010; Goodfellow et al. 2016). Conditional entropy values are used to evaluate the mathematical relationship of one variable set to another variable set and entropy values indicate the suitability for plotting one formant against the formant, for example, F1 vs. F2 or F1 vs. F3 plots.

The conditional entropy values were calculated with a modified version of Mutual Info 0.9 cross-platform program package (Peng 2002) operating in Matlab ™. Conditional entropy calculations provided numerical data that shows the relationship of formant values to: F1 to F2, F1 to F3, and F2 to F3.

Materials and methods

This experiment was designed to compare the entropy values of the vowel portions of target words that were produced in clearly spoken sentences. This experiment used words that were produced for a previous perception-production study that compared English and Greek vowel spaces (Botinis, et al., 2022). Male and a female native New Mexico English speakers, using carrier sentences, recorded the words: *bit, beat, bet, bat, boot, butt, bought,* and *bot*, producing the monothongs /iː, ɪ, e, æ, uː, ʌ, ɔ, ɑː/, which represent the corner vowels and edges of their respective dialect vowel space. Speech recordings were produce using a Røde N microphone in a GretchKen™ Industries acoustic sound booth. The sentences were spoken multiple times clearly and the key words in the sentence were spoken with brief silence before and after each word. Formants were next created using PRAAT, (standard settings of 5,500 Hertz ceiling and 5 formants) and the formant data was recorded in a spreadsheet. Entropy calculations were completed for natural speech and the data was displayed in a self-populating datasheet for comparisons.

The Speechify ™ cloning program was trained on the New Mexico English carrier sentences. Speechify™ was used then to produce the same cloned sentences. The cloned vowel formants were determined again using PRAAT, and the values were copied into data sheets. Entropy calculations were completed for the cloned speech and the data was displayed in a self-populating datasheet for comparisons.

Results

The vowels spoken by the female participant (/iː, ɪ, e, æ, uː, ʌ, ɔ, ɑː) were examined and it was found that in the eight vowels, at least one formant was dominant and at least one formant was dependent. This is a common pattern observed in our numerous entropy calculations for natural speech. The dominant formant would be best plotted on the x-axis and the dependent formant would be best plotted on the y-axis.

For the cloned speech samples of the female speaker, the entropy values of each vowel portion have significantly altered dominant formants, compared to the formants of natural speech. Entropy relationships between cloned formants are significantly weaker than the entropy values of natural speech formants. This also indicates that plotting cloned formants would not produce reliable graphs.

The vowels spoken by the male participant (/iː, ɪ, e, æ, uː, ʌ, ɔ, ɑː) were examined and it was found that in the six vowel pairings, at least one formant was dominant and at least one formant was dependent. Again, the dominant formant would be best plotted on the x-axis and the dependent formant would be best plotted on the y-axis.

For the male speaker cloned speech, the entropy values of the target vowels from the words *bat, beat, bet*, and *bit* showed that each cloned vowel formant has altered dominant formants. The conditional entropy values increased for cloned speech, indicating a weakening of the relationship between formants. For example, the cloned vowels derived from words *boot* and *bot* showed an entropy increase in formant relationships, weakening and changing F3 to the dominant formant. The cloned vowels derived from the words *bought* and *butt* showed reversed dominant formants with the dependent formants, compared to natural speech.

Table 1. Conditional entropy differences between speech and cloned speech for the female speaker.

Female speaker	bat /æ/	beat /ɪ/	bet /e/	bit /i:/	boot /u/	bot /ɑ:/	bought /ɔ/	butt /ʌ/
F1-F2 conditional	0.068966	0.095238	0.083333	0	0.095238	0.181818	0.442684	0
F2-F1 cond	0.137931	0	0	0	0.095238	0.090909	0.076923	0.1
F1-F3 cond	0.137931	0	0	0	0	0	0.076923	0
F3-F1 cond	0.137931	0	0	0	0.095238	0.090909	0.076923	0.1
F2-F3 cond	0.137931	0	0	0	0	0	0.076923	0
F3-F2 cond	0.068966	0.095238	0.083333	0	0.095238	0.181818	0.442684	0
Cloned words	**/æ/**	**/ɪ/**	**/e/**	**/i:/**	**/u/**	**/ɑ:/**	**/ɔ/**	**/ʌ/**
F1-F2 conditional	0.295385	0.094118	0.179775	0.144928	0.1	0.133333	0.294872	0.094118
F2-F1 cond	0.306119	0.806227	0.384996	0.659562	0.673429	0.133333	0.557818	0.361822
F1-F3 cond	0.178921	0.164706	0.269663	0.144928	0.184436	0.222222	0.34785	0.188235
F3-F1 cond	0.32336	0.806227	0.384996	0.659562	0.673429	0.133333	0.557818	0.361822
F2-F3 cond	0.178921	0.164706	0.269663	0.144928	0.184436	0.222222	0.34785	0.188235
F3-F2 cond	0.312626	0.094118	0.179775	0.144928	0.1	0.133333	0.294872	0.094118

Table 2. Conditional entropy differences between speech and cloned speech for the male speaker.

Male speaker	bat /æ/	beat /ɪ/	bet /e/	bit /i:/	boot /u/	bot /ɑ:/	bought /ɔ/	butt /ʌ/
F1-F2 conditional	0	0	0.321661	0.26087	0	0	0.133333	0.1
F2-F1 cond	0.173913	0.222222	0.190476	0.173913	0.086957	0.086957	0	0
F1-F3 cond	0	0.222222	0	0	0.347826	0.347826	0.066667	0.1
F3-F1 cond	0.173913	0.222222	0.190476	0.173913	0.086957	0.086957	0	0
F2-F3 cond	0	0.222222	0	0	0.347826	0.347826	0.066667	0.1
F3-F2 cond	0	0	0.321661	0.26087	0	0	0.133333	0.1
Cloned words	**/æ/**	**/ɪ/**	**/e/**	**/i:/**	**/u/**	**/ɑ:/**	**/ɔ/**	**/ʌ/**
F1-F2 conditional	0.153846	0.210526	0.173913	0.111111	0	0	0	0
F2-F1 cond	0.076923	0.210526	0.347826	0.222222	0.08	0.166667	0.125	0.222222
F1-F3 cond	0.076923	0.105263	0.173913	0.111111	0.16	0.166667	0.27359	0
F3-F1 cond	0.076923	0.210526	0.347826	0.222222	0.08	0.166667	0.125	0.222222
F2-F3 cond	0.076923	0.105263	0.173913	0.111111	0.16	0.166667	0.27359	0
F3-F2 cond	0.153846	0.210526	0.173913	0.111111	0	0	0	0

Conclusions

Conditional entropy has been used to detect patterns in data by providing numerical values that show their relative degree of overlapping information, showing how dependent one formant is on another.

The clones of the target words were perceptually similar to the original spoken words, but the entropy values for the clones of showed considerable changes. Generally, conditional entropy values were higher for cloned speech, which is an indication of increased randomness in the sound signal. Higher clone entropy values indicate weaker relationships between the formants for the cloned speech, which would translate to weaker or less clearly defined harmonics.

This study is limited, and it is exploratory, but strong positive results indicate that the vowel formants undergo entropy and often formant priority changes when they are cloned. More testing will be used to duplicate the method we have described using a variety of speech samples as we look for predictable patterns in the cloned speech entropy values.

We theorize that identifying entropy changes due to cloning may contribute to linguistic forensic examinations. Cloned speech has the potential to be used fraudulently and identifying cloned speech by its entropy values should be a way to determine the authenticity of speech.

References

Botinis, A., Goertz, G., Kontostavlaki, A., Anderson, T. 2023. Vowel discrimination of American English. ExLing 2023 Athens: Proceedings 14th International Conference of Experimental Linguistics, October 13-16. Athens, Greece.

The MathWorks Inc. 2023. MATLAB version: 9.13.0 (R2022b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com.

Goodfellow, I., Bengio, Y., Courville, A. 2016. Deep Learning. Cambridge, Massachusetts: The MIT Press.

Haglund, J.F. Jeppsson, H. Strömdahl. 2010. Different Senses of Entropy – Implications for Education. Entropy 12, 490-515. Doi: 10.3390/e12030490.

Peng, H. 2022. Mutual Information computation. (www.mathworks.com/matlabcentral/fileexchange/14888-mutual-information-computation),