21 Neural audio captioning

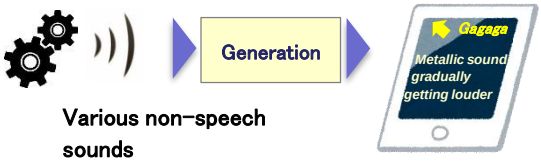
- Generating text describing non-speech audio -

Abstract

Recently, detection and classification of various sounds has attracted many researchers attention. We propose an audio captioning system that can describe various non-speech audio signals in the form of natural language. Most existing audio captioning systems have mainly focused on "what the individual sound is," or classifying sounds to find object labels or types. In contrast, the proposed system generates (1) an onomatopoeia, i.e. a verbal simulation of non-speech sounds, and (2) an sentence describing sounds, given an audio signal as an input. This allows the description to include more information, such as how the sound sounds and how the tone or volume changes over time. Our approach also enables directly measuring the distance between a sentence and an audio sample. The potential applications include sound effect search systems that can accept detailed sentence queries, audio captioning systems for videos, and AI systems that can hear and represent sounds as humans do.

Detailed audio captioning

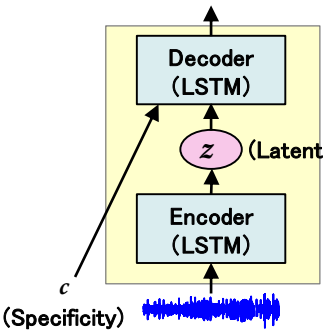
・Given an audio input, the system describes not only what the sound source is, but also how it is sounding and how it is changing over time, as a natural language sentence. NEW

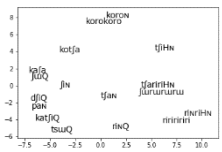


Various non-speech sounds

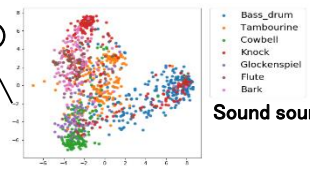
Method

A high-pitched fricative noise is …





Onomatopoeias mapped in the latent space



Sound sources

Instrumental sounds mapped in the latent space

1. Onomatopoeic mode: output a word simulating the sound
2. Sentence mode: output a sentence describing the sound

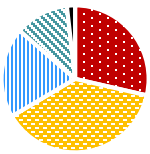
【Point】No unique correct answer for captioning
→【Proposal】 Can control the degree of detail by conditioning the decoder by "Specificity" input
Specificity: Sum of the amount of information contained in the output text.

Experimental results

The description matches the sound:

Appropriateness of the output sentence (without specificity conditioning)

With the specificity control, the sentence is:



Effectiveness of the specificity conditioning

Examples of description for a base drum sound (English translation from Japanese)

c	Generated sentence
–	A low sound rings for a moment
20	A low sound sounds for a moment
50	A low, striking sound sounds as if something is dashed on a mat
80	A very low-pitched drum is played only once
110	A faint, low-pitched sound sounds as if something is hit dully, and it soon disappears

References

[1] Shota Ikawa, Kunio Kashino, "Generating sound words from audio signals of acoustic events with sequence-to-sequence model," In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), April 2018.

[2] Shota Ikawa, Kunio Kashino, "Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds," In Proc. Detection and Classification of Acoustic Scenes and Events (DCASE 2018), November 2018.

Contact

Kunio Kashino Email: cs-liaison-ml at hco.ntt.co.jp
Media Information Laboratory

Innovative R&D by NTT
Open House 2019