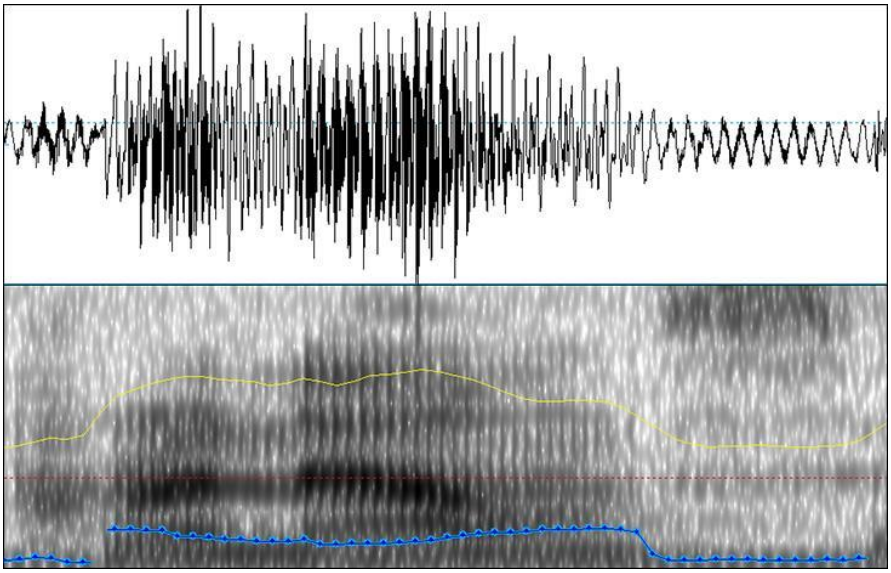
# Was That a Question? Neural Networks Need to Know

*By* Martin Anderson



*First published* ***June 14th, 2017*** *at:*
*https://www.intelligentautomation.network/artificial-intelligence/articles/was-that-a-question-neural-networks-need-to-know*
*Web-archived version*

**As far as Machine Learning is concerned, it's definitely the way that you say it.**

One of the key challenges on the road to perfect AI-driven speech recognition is the development of reliable frameworks that can understand not just what we are saying, but how we are saying it. Without this faculty machine-driven language processing will never evolve beyond accurate transcription into a genuine understanding of intent.

The challenge is tough enough in English, where emphasis can completely alter the meaning of a sentence:

*I* didn't steal that. (Someone else stole it)
I *didn't* steal that. (I negate the allegation that I stole it)
I didn't *steal* that. (I own it, theft does not apply)
I didn't steal *that.* (But I did steal something else)

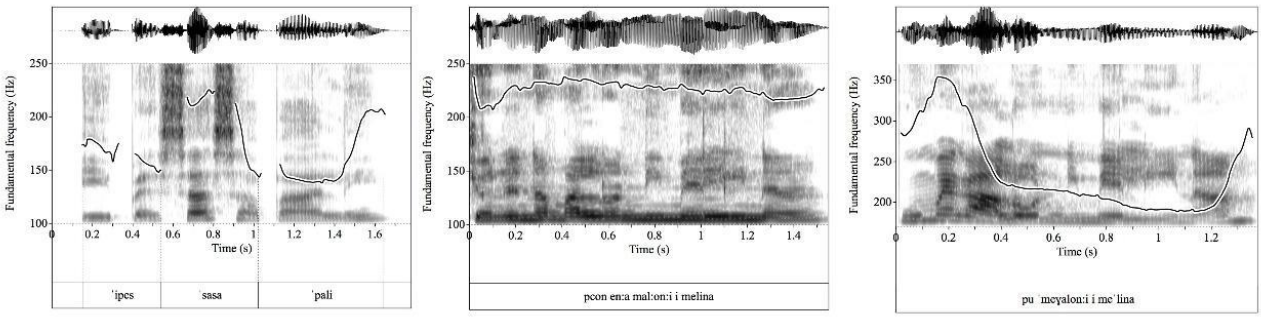
In Mandarin, vocal tone can completely change the meaning of a word, and recognizing prosody becomes critical. The word 'Ma' means 'mother' when voiced at a constant high pitch, and 'horse' when said with a low and then rising pitch. This example represents a particularly unfortunate re-use of syllables, emphasizing the importance of tonal interpretation in machine translation systems.

## A Visual Approach to Sound

Machine Learning uses diverse approaches to the creation of autonomous and supervised Neural Network-based speech recognition and translation systems. The two vanguard approaches in this period are Long Short-Term Memory (LTSM) and Convolutional Neural Networks (CNN).

New work by Jean-Philippe Bernandy, of the University of Gothenburg in Sweden, has found that despite the traditional boundaries between applicability in RNNs and CNNs (RNNs for speech analysis, CNNs for image processing), using a Recurrent Convolutional Network can be notably more efficient in analyzing speech intonation, demonstrating 95 per cent effectiveness vs 82 per cent for CNNs.

It's a development which emphasizes why image recognition is becoming such a pivotal facet of AI research. Even where the material is sound-based or in some other way abstract and non-visual, ultimately most quantifiable data resolves to a graph.




Bernandy's intonational model used 1966 voice recordings by 25 female speakers and 2860 additional question recordings from 20 female speakers of Cypriot Greek.

The first recordings were modelled from the phrase 'You said CVCV again' (C = consonant, V = vowel), while the second consisted of questions of variable length.

The LTSM network, expected to score higher in speech-based tasks, received an 82 per cent accuracy score, while the vision-based Convolutional Neural Network managed 95 per cent accuracy.

Bernandy notes also that the CNN achieved these results without manually tagged data regarding accent pitches. On a more cautionary note, he observes that the data is very regularized by nature of the sampling and the participants, and that it remains to be seen if similar results could be obtained on a more varied dataset.

## The Question of Tone

For the purposes of security applications the development of indices for mood analysis in voice recording is one obvious imperative in creating Machine Learning systems that take tone into account.

There are additional possibilities for use of tonal voice data in medical research. In 2016 a group of Australian researchers developed new technologies for mood-discernment using a development of Deep Neural Networks (DNN - see glossary) called a Deep Belief Network. The technique compares the tone of the user's voice against a baseline provided by the median tone of the conversation group; in practice, this could potentially identify one 'unhappy' person in a group of happier ones.

A further challenge for automated or context-aware tonal recognition is the rising inflection, aka 'upspeak', which can render the (already considerable) problem of identifying a rhetorical question quite chronic in certain locales. The inverse of the upspeak problem is 'vocal fry' , where the speaker maintains an artificially deep tone for dramatic effect.

It seems that tone recognition algorithms will eventually need a fairly elaborate schema of potential psychological explanations for these anomalies, and will also have to take cultural and geographical factors into account regarding them.

## Taking Voice Research on the Road

Truly widespread data-gathering applications capable of generating (hopefully with user consent) a large corpus of voice data will either need to operate within the processing and memory limitations of mobile devices (locally intensive) or to maintain unusual levels of cloud connectivity (bandwidth intensive). Most research work in this field is currently waiting on a solution to the conundrum, and remains laboratory-bound.

One research group from the Institution of Automation at the Chinese Academy of sciences has presented a model for a quantized CNN, whilst MIT presented a new chip prototype in 2016 intended to enable RNN functionality via energy-saving optimizations. Dubbed Eyeriss, the chip reduces the number of intra-core exchanges on a GPU.