Challenges and solutions pertinent to machine learning-based audio recognition

Sijie Yang *

Beijing National Day School, Beijing China

* Corresponding Author Email: 18911270626@163.com

Abstract. The problem of transcribing data from an acoustic waveform has yielded multiple approaches over the last decades. The current preferred approach involves a three-stage model that breaks the problem into its constituent stages, each with an equivalent model. The first stage divides pure acoustics and language study using a Bayesian model. The second stage focuses on the acoustics model; this model makes the most sufficient and efficient division. The third stage focuses on the acoustics model; this model provides complete instruction on how to compute the probability required by the first stage, given the division specification declared within the second stage.

Keywords: Audio to text; speech to text.

1. Theoretical assumptions

We assume that the audio we receive as input is in a digital format, without intense noise or bad recording quality, and is spoken by a regular human with normal pitch and speed.

The assumption is plausible because any smart phone would fulfill such requirements, and the need of audio to text often happens in online videos, rather than music industry, where variable pitch and mumbling do exist.

2. Symbol specifications

Table 1. Standards.

Noun	Symbol	Explanation
Audio	X	An Audio represented in discrete wave form.
Identified Text	Y*	The text with most probability, chosen as the identified text.
Text	Y	Any Text, represented by words.
State Sequence	S	A sequence of states (The smallest division unit), converted from a sentence (A sequence of words).
Alignment precision	φ	The number of uniform divisions to perform on the audio, as well as the length of the aligned sequences.
Aligned sequence	H	An alignment of all the alignments of the original sequence S, to length φ.

3. Stage 1: Acoustics-language division

$$Y \in YGG \tag{1}$$

To find the most credible identification of an Audio, we can iterate through all possible texts, represented by the set YGG.

$$\begin{aligned} Y^* &= \arg\max_Y P(Y \mid X) \\ &= \arg\max_Y \frac{P(X \mid Y)P(Y)}{P(X)} \\ &= \arg\max_Y P(X \mid Y)P(Y) \end{aligned} \tag{2}$$

$$\begin{aligned} AcousticModel(X,Y) &\equiv P(X \mid Y) \\ LanguageModel(Y) &\equiv P(Y) \end{aligned} \tag{3}$$

From a Bayesian perspective, the probability of text Y, given Audio X, is equivalent to that of Audio given text multiplied by the probability of text Y, divided by the constant P(X), which does not affect maximizing the argument. We give the first part the name “acoustic model” and the second part the name "language model”. The acoustic model focuses on whether we were to say a given text out loud and how likely the audio will be like the one given. Language focuses on how meaningful or likely the sentence will occur in real life, which includes grammar, rationality, and morality as its variables. For instance, the sentence “poecillia verb jump walk cannibalism is good go up down left right” would have a lower output probability from the language model because of its bad grammar and moral context.

4. Stage 2: Acoustic features subdivision

We must make more detailed subdivisions to describe audio using limited, fixed acoustic features rather than words. The first layer of division is the phoneme, a combination of actions of our lips, tongue, teeth, jaws, vocal cords, and nose. This is the minimal unit in language definition and oral action but not in waveform definition. A phoneme’s actual waveform representation is heavily affected by its pre-and post-phonemes because our tongue, teeth, and jaws have a finite speed and are certainly unable to make instant, step function-like transitions; even if we do so, the air-particles have a finite time to respond to our action. With this in mind, we define a subdivision unit inherited from phoneme. The new unit assumes that the waveform representation is solely affected by the one before and after. Therefore, the new unit was defined by three phonemes, given the name tri-phoneme.

From the statistical results collected, we find that units share more common features if we further divide a tri-phone chronologically into three segments. The segments are given the name “State”.

$$\begin{aligned} &Word \in Sentence(Y) \\ &Word \xrightarrow{divide} Phoneme \xrightarrow{convolution} Tri-phone \xrightarrow{divide} State \\ &State \in StateSequence(S) \end{aligned} \tag{4}$$

With all the subdivisions, we can implement an algorithm that represents a sentence in a chronological sequence of states.

$$S = subdivide(Y) \tag{5}$$

5. Stage 3: Alignment and identification

$$P(X \mid Y) = P(X \mid S) \tag{6}$$

We now have an equivalent expression state sequence S of sentence Y; the only thing left to do is to find the likelihood of the audio, given the sequence of states. Once again, we divide the sequence into individual components. For example, we may ask, given a single state, how likely it is to be a segment of audio x? We define that as the “Emission Probability”. To obtain that probability, we first have to extract features from the messy waveform that no human can read (data in the form of the waveform is just incomprehensible). The classic approach to analyzing audio is to convert it from a time field to a frequency field by a Fourier transformation. This concerns the degree to which low, medium, and high frequencies exist in an audio sample. We then selected critical frequencies related closely to human pronunciation to enhance the data.

The frequency distribution is then flattened to a vector, therefore able to be put into an N-dimensional Gaussian distribution inference and obtain a likelihood value as the probability output. This method is known as the Gaussian Mixture Model (GMM).

$$\begin{aligned} &a \in State \\ &P(x \mid a) = GMM(x, a) \end{aligned} \tag{7}$$

Now that the individual logic has been solved, we can consider the sequence as a whole.

Similarly, the sequence's chronological nature means we can implement a state transition probability model. The terminology means that we have a probability chart recording how likely it is to change from one state to another so we can determine the likelihood of a sequence by iterating and multiplying each transition that occurs in the sequence.

As we assume each transition solely depends on the current state, the transitions can be iterated in random order or even sampled randomly as long as the probability expectation is the same.

$$\begin{aligned} &s_i \in S \\ &n = length(S) \\ &Credible(S) = \prod_{i=2}^{n} P(s_i \mid s_{i-1}) \end{aligned} \tag{8}$$

Of course, we can simply assume that any state would have a fixed length of time. Therefore, each state should accumulatively chronologically correspond to a segment of audio signal having its own acoustics feature. With this in mind, we can derive a formula of P (X|S) in the form below:

$$\begin{aligned} &X_{divided} = razer(X, S) \\ &x_i \in \mathrm{X}_{divided} \\ &n = length(S) = length(X_{divided}) \\ &P(X \mid S) = Credible(S) \prod_{i=1}^{n} P(x_i \mid s_i) \end{aligned} \tag{9}$$

This causes certain problems to arise, however. What if the time length of X does not perfectly match with the time sequence provided by S? I have assumed that the audio was provided as normal speed, but people certainly do have slight variations in the speed of each word and the intervals between. By applying such a simple approach, we have lost most of our universality and applicability.

A solution to this problem exists. The credible algorithm is preserved, as well as the individual likelihood between an audio segment and a state, while the sequence of states is modified. We define an alignment precision bigger than n (the original length of state sequence S), denoted by symbol φ. theoretically we then search for all possible arrangements of S with a total length of φ. The likelihood between the audio and a possible arrangement was calculated as treating the arrangement as the original sequence S.

With this in mind, the likelihood between the original sequence S and X is approximated by the sum of the likelihood between each arrangement (denoted by H) and X.

$$\begin{aligned} &X_{grid} = uniformGrid(X, \Phi) \\ &\Phi = length(h) = length(X_{grid}) \\ &x_i \in X_{grid} \\ &h_i \in H \\ &P(X \mid H) = Credible(H) \prod_{i=1}^{\Phi} P(x_i \mid h_i) \\ &P(X \mid S) \approx \sum_{H \in Alignments(S, \Phi)} P(X \mid H) \end{aligned} \tag{10}$$

6. Model review

6.1. Advantages of the model

This model is revolutionary compared to a single word recognition system based on a single neural network, as it's no longer superficial, but instead have real world applications that leads to productivity boost like video subtitle generation.

This model, as a framework, is able to become a state-of-the-art approach, as long as the correct algorithm is used for its module. [1]

6.2. The innovations of the model

The main innovation of this model is the use of a Bayesian perspective. It sub-divides the model and breaks down the question into two subjects for which a priori knowledge is available.

The other innovation of this model is the modularized design in the process, which allows different algorithms like GMM or DNN to be chosen as the implementation to a single function.

7. Conclusion

The presented model is able to make use of prior knowledge available, leading to a more efficient workflow. The model is amenable to computers with low computing power, and when implemented along with a translation model, can give every foreign language video a subtitle that people can understand, among other practical applications.

Reference

[1] Feng Jiajun (2016) Research and System Implementation of Violent Audio Scene Classification Technology (Doctor Dissertation, Harbin Institute of Technology).

[2] Zhang Xiaolong, &Peng Yi (2023) Audio recognition method based on residual network and random forest The 6th China Computer Society Big Data Academic Conference.

[3] http://www.bnext.com.tw/article/41414/bn-2016-10-19-020437-216 [Yu, et al., INTERSPEECH'16].