INTERSPEECH 2020: Show & Tell Contribution
October 25–29, 2020, Shanghai, China

Soapbox Labs Verification Platform for child speech

Amelia C. Kelly, Eleni Karamichali, Armin Saeb, Karel Veselý, Nicholas Parslow, Agape Deng, Arnaud Letondor, Robert O'Regan, Qiru Zhou

SoapBox Labs, Dublin, Ireland
amelia@soapboxlabs.com

Abstract

SoapBox Labs' child speech verification platform is a service designed specifically for identifying keywords and phrases in children's speech. Given an audio file containing children's speech and one or more target keywords or phrases, the system will return the confidence score of recognition for the word(s) or phrase(s) within the the audio file. The confidence scores are provided at utterance level, word level and phoneme level. The service is available online through an cloud API service, or offline on Android and iOS. The platform is accurate for child speech from children as young as 3, and is robust to noisy environments. In this demonstration we show how to access the online API and give some examples of common use cases in literacy and language learning, gaming and robotics.
Index Terms: speech recognition, child speech, educational technology, pronunciation assessment, child literacy

1. Introduction

The SoapBox Labs child speech verification and pronunciation assessment technology platform (SoapBox Verification) is a service for that allows developers and content providers to voice-enable their products services. It can be conveniently accessed by web API for customers and is also available offline on Android and iOS platforms. The system is built on models trained using many hours of child speech from real-world situations. We leverage the data and models as well as a verification algorithm to create accurate speech verification specifically for children's voices using state-of-the-art deep learning algorithms. While SoapBox Labs offer separate services for fluency assessment (SoapBox Fluency) and automatic speech recognition (SoapBox Speech-To-Text), there are many applications for which a verification solution provides the best functionality and versatility. SoapBox Verification has been built for those who are developing literacy aids and language learning products, multiple choice games, or dialogue systems. These applications have the following things in common and present the some of the best applications suited for verification:

1. the developer knows in advance the word(s) or phrase(s) they want to score e.g. child reads word/sentence, has been asked a question, voice commands to control game/toy/IoT devices etc.
2. there is a maximum of 100 words/phrases of interest (these can vary on-the-fly – the API accepts next-to-unlimited vocabulary)
3. a score is required for each of these particular words or phrases, either at the utterance, word or phoneme level

Unlike traditional speech-to-text systems (like Soapbox STT) our verification algorithm only returns information related to the words or phrases specified and will not substitute for more likely output. This allows the developer to focus on the pronunciation score relevant to their application. The verification system does not require tuning to particular domains and can work for any input word or phrase – the user can request a pronunciation score for any vocabulary. A common strategy that has been attempted for keyword spotting in speech has been to use automatic speech recognition (ASR). However, verification provides greater accuracy and flexibility for many of the same use-cases. When using speech recognition for keyword spotting, it is possible and probable that the transcription will not contain all of the words the child has actually said. In this case, it is impossible to calculate confidence scores for every word, prompted or otherwise. In contrast, supplying the keyword or phrase targets with the audio file, as with the verification system, simultaneously allows for a pronunciation score to be returned for every word regardless of how unlikely it was that the word was spoken by the child. SoapBox Verification does not require tuning to a particular domain and can work for any input word or phrase. While the ASR approach (like that offered by SoapBox Speech-To-Text and SoapBox Fluency) is very useful for longer passages or audio search for example, verification can provide similar functionality with a smaller footprint and more versatility.

2. Directions for use

SoapBox Verification is officially released and is generally available as a RESTful Web Service. Once authenticated, speech verification requests should be sent via HTTPS to: https://api.soapboxlabs.com/v1/speech/verification The parameters that should be included for a successful response are given in Table 1. An HTTP request can be sent to the URL containing an audio file (min 16 kHz, 16 bit PCM, wav format) and one or more verification targets.

```
curl -H "X-App-Key:YOUR_API_KEY_HERE"
-F "file=@AudioFile.wav"
-F "category=right"
-F "user_token=abc123"
https://api.soapboxlabs.com/
      v1/speech/verification
```

Targets are specific words or phrases that you want to search for within an audio file. Targets can also be keywords that you might want to search for within a larger phrase or sentence. Every verification request requires the presence of at least one target. Multiple targets may also be specified. If an audio file is

Table 1: *Description of fields in CURL request*

file	audio file to be analysed
category	targets to be checked for in audio
user_token	unique id that represents the speaker

Table 2: *Description of fields in JSON results object*

user_id	user_id specified in request
language_code	Language spoken e.g. en_GB
result_id	A unique identifier for the request
time	UTC time request was processed
results	results for targets specified
category	target(s) specified in request
hypothesis_score	overall score for target
word_breakdown	results for tokens in target phrase
word	each word in the target
target_transcription	phonetic transcription of word
quality_score	score for word/phoneme
phone_breakdown	phonetic breakdown of each word
phone	constituent phone of word

uploaded with one target, a response will be returned containing the confidence score for that target. Depending on the application, this can be interpreted as a likelihood that the given word or phrase was spotted within the audio file, or as a measure of how well the word or phrase was pronounced. If an audio file is uploaded with multiple targets, the response will contain the confidence scores for each target. The most common application of multiple targets is "command and control" or "multiple choice" use cases, such as voice-activated action games, and conversational or dialog games. Once the verification engine has processed the supplied audio file and targets, a JSON result object will be returned. This is an example of a typical JSON response. The fields are described in Table 2.


```
{
  "user_id": "abc123",
  "results": [{
    "category": "left",
    "hypothesis_score": 26.0,
    "word_breakdown": [{
      "word": "left",
      "target_transcription": "l eh f t",
      "quality_score": 26.0,
      "phone_breakdown": [{
        "phone": "l",
        "quality_score": 20.0
      }, {
        "phone": "eh",
        "quality_score": 5.0
      }, {
        "phone": "f",
        "quality_score": 6.0
      }, {
        "phone": "t",
        "quality_score": 98.0
      }]
    }]
  }],
  "language_code": "en-GB",
  "result_id": "abc123-1_1571927942711",
  "time": "1970-01-01T14:39:02.988Z"
}
```


3. Demonstration of SoapBox Verification

Many use-cases the developer knows in advance what words and phrases are of interest. These are the use-cases that SoapBox Verification are best suited for. In this demonstration we show how the verification system can be used in these various use-cases.

3.1. Literacy and language learning: pronunciation

Verification can be used in literacy and language learning applications, where a child is shown a word or phrase and the scores returned at the utterance, word and phoneme level can provide a measurement of pronunciation assessment to the student, parent or educator. In this demonstration the child is prompted for the phrase "the horse runs fast" but the child actually says "the horse **run** fast". The verification system returned a low score for the "/s/" phoneme in "runs" because it was not spoken. This is illustrated in Figure 1. Similarly when the child is prompted to read "This is an arachnid" they leave out the word "an", and this is correctly identified by the Verification system.

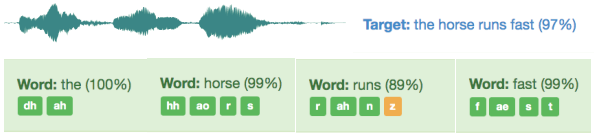

Figure 1: *Pronunciation assessment in literacy application*

3.2. Conversational platforms: multiple target search

Verification can also be used in scenarios where the child might have an open-ended response and the user may only be interested in certain parts of the response. Take for example a comprehension test – the child has read a passage about a dog and we want to check how much they understood by asking the to describe the dog. Many answers are acceptable, and verification allows to simultaneously search for different acceptable words and phrases. Similarly, in Figure 2 the child is asked to describe what they see in a picture. The child said "chicken", so both *chick* and *chicken* score highly.

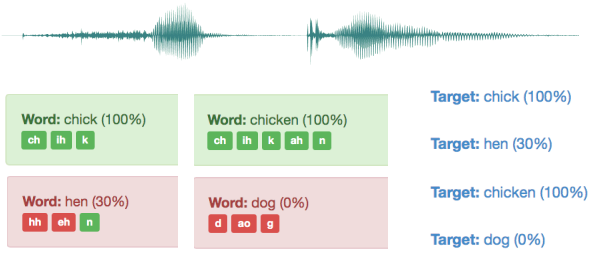

Figure 2: *Multiple targets in conversational application*

4. Conclusions

We presented the SoapBox Labs child speech verification platform, a service designed specifically for identifying keywords and phrases in children's speech. We provided directions on how to use the cloud API and demonstrated how the results returned could be used in various applications for children. We demonstrate these use cases in the accompanying video. The accuracy and versatility of the SoapBox Labs verification platforms, as well as its robustness to noise allows it to add real benefit to ed-tech, gaming, robotics platforms, and any other platforms where it's essential to accurately decipher child speech.