*INTERSPEECH 2016: Show & Tell Contribution*
September 8–12, 2016, San Francisco, USA

# English Language Speech Assistant

*Xavier Anguera, Vu Van*

ELSA Corp.
{xavier, vu}@elsanow.io

## Abstract

This show&tell demo presentation showcases ELSA Speak, an app for English Language pronunciation and intonation improvement that uses speech technology to assess the users speech and to offer consistent feedback on the errors the students make.

**Index Terms**: pronunciation feedback, L2 learning

## 1. Introduction

This demo falls within the area of Computer Assisted Language Learning (CALL) [1]. CALL has gained a lot of interest lately, mostly due to the advances in speech recognition that allow students to get better understood by the computer and which widened the possibilities to automatically evaluate the students' voice [3, 7].

In the proposed demo we focus on pronunciation and intonation improvement [2, 4, 5, 6]. Pronunciation and intonation are the hardest skills to master in language learning, because these skills require individual attention, repetition and precision. 1:1 pronunciation training with speech therapists are too expensive and not scalable. We observed that users tend to resort to YouTube or TV show to mimic American accents, but that is a one-way learning and they do not usually have anybody to get feedback from. Linguist experts point out that the fastest way for language learners to improve their pronunciation is to receive detailed feedback on their particular errors and phonetic hints to fix those errors.

To verify this user need we did a customer survey in early 2015 with 2,000 English language learners and 90% of them indicated they need most help with pronunciation. Researches have shown people who speak English with accents are perceived to be 30% less trustworthy, and indicates a 40% income gap between those who speak English well and those who don't.

To mitigate this problem we have built a robust and scalable system that is currently serving thousands of users daily, and an app that is available both in IOS and Android platforms. We build our speech recognition technology to detect pronunciation errors at phoneme level as people speak English, and offer instant feedback to fix them. Our vision is to enable 1.5 billion language learners to speak English well and be better understood, and unlock new opportunities.

## 2. ELSA application

The first product we have launched is a mobile app called ELSA Speak, which allows users to practice and improve their pronunciation and intonation skills through a set of exercises that are evaluated on our servers. We have so far developed Android (available since November 2015) and Apple IOS (available since March 2016) versions of the app and we are constantly updating them following feedback from our users.

The app currently offers three main exercise types: pronunciation, intonation and conversation training. These are described next and illustrated in Figure 1.

**Pronunciation exercise** Users speak the proposed word or phrase and get the feedback (with a color code) for each phoneme, as well as phonetic hints to fix existing errors. See Fig. 1b.

**Intonation exercise** Users practice word syllable stress as well as sentence intonation and rhythm. See Fig. 1c.

**Conversation exercise** Users immerse in practicing real-life conversations and receive instant feedback on their pronunciation and intonation at word level. See Fig. 1d.

In addition, we have a free-text input mode where users can listen to the sample pronunciation of any word or sentence and then practice it and get feedback instantly on how it should be pronounced.

## 3. ELSA System Architecture

The system architecture powering the ELSA Speak app implements a client-server processing scheme. For every trial the app submits the spoken audio to the server as well as information about the user and the exercise being spoken. It then gathers the processed results and presents them to the user. Audio is streamed to the server in real time so that processing in the server can start before the user finishes speaking the sentences, therefore receiving a quicker answer.

When speaking a sentence, the user is instructed when to start speaking (with a beeping sound) and then endpointing detection is performed on the server to stop processing and return results.

In the server we are using the Kaldi [1] as speech processing engine with custom trained DNN models. In order to ensure scalability of the service we built all necessary components into a single quad-core machine and replicated them using Amazon's elastic load balancing capabilities, with computational nodes in multiple regions, selected at run time via DNS resolution. Each note has all L1/L2 language pairs available at runtime so that any user trial can be processed by any machine.

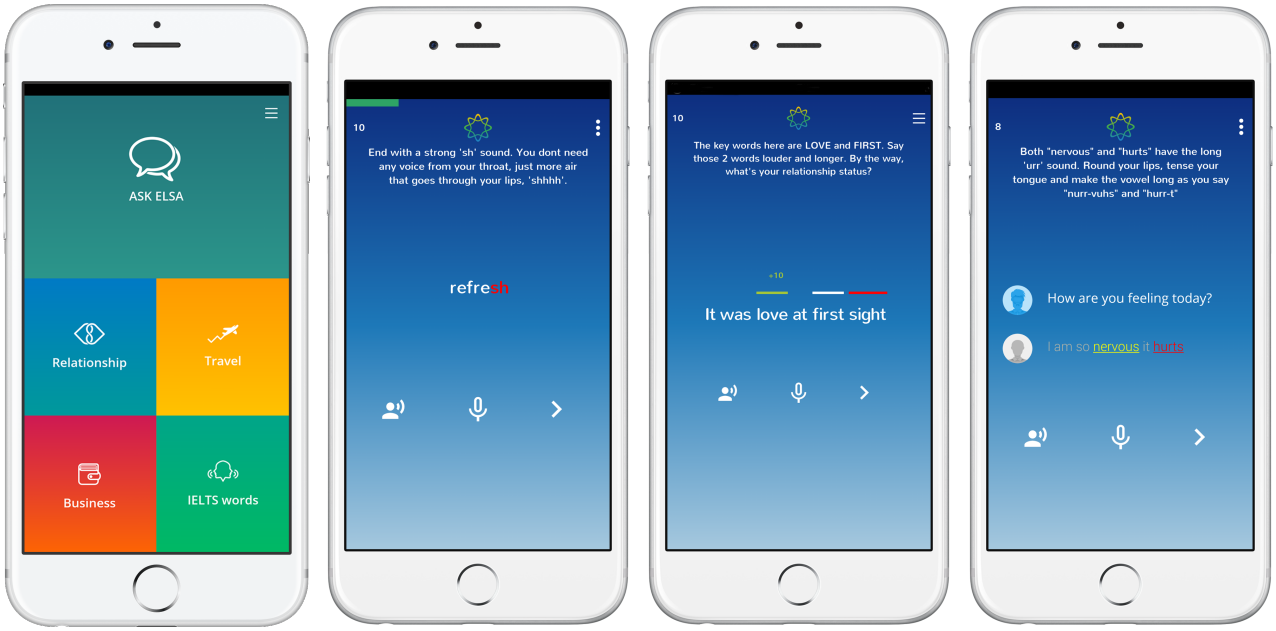
### 3.1. Acoustic Model Training

The ELSA Speak app is available for use by anyone by using our generic English acoustic models. In addition we are also developing specific acoustic models for users of a native language L1 that wish to learn a second language L2 (English for now). These models are able to pinpoint the most common problems that a speaker of L1 will face when speaking L2, so that more specific feedback can be given to the user.

## 4. App evaluation

Since the launch of the app we have had many users sign up and use the app on a regular basis. So far (as of March 30th, 2015) we have processed over 3 million user trials in our servers for all of our exercise types.

[1] http://kaldi-asr.org




(a) A Main menu (b) B Pronunciation exercise (c) C Intonation exercise (d) D Conversation exercise

Figure 1: Example screenshots from the ELSA app

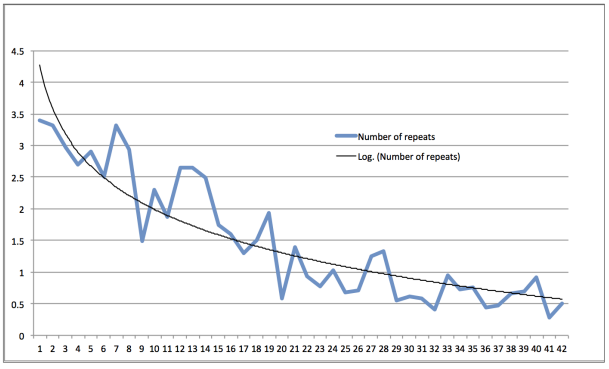



Figure 2: *Average number of repetitions to get a word right as a function of number of trials.*

During the initial phases of the product we performed many interviews with users to define the set of features that they would value most and modified the product accordingly. In addition, we continuously collect feedback from users to improve the app in future versions. To illustrate how the app can help improve intonation and pronunciation skills we have analyzed the data from several regular users of the app. Figure 2 shows the average number of repetitions that a group of 50 regular users needed to perform to get a word right (no errors), as a function of the total number of words they spoke. We can see how our users improve as they use the app. In order to discard the possibility of users learning to trick the app, Figure 3 shows the relative nativeness improvement (we define nativeness as a function of the errors and warnings a user gets over time) for one of the previous users as a function of how many words she exercised.

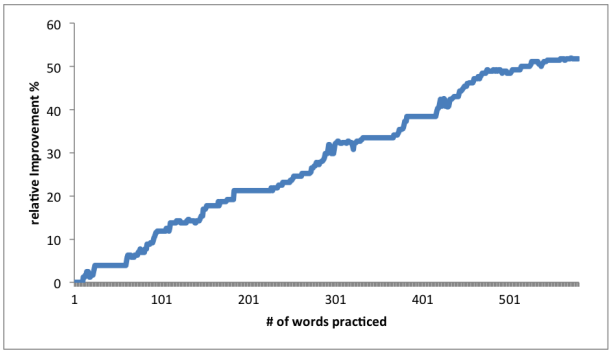



Figure 3: *Relative nativeness improvement as a function of the number of trials.*

## 5. References

[1] M. Levy, "Computer-assisted language learning: Context and conceptualization," *Oxford University Press.*, 1997

[2] G. Kawai and K. Hirose, "A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training," *ICSLP*, 1998

[3] J. Dalby and D. Kewley-Port, "Explicit pronunciation training using automatic speech recognition technology," *Calico Journal*, 425-445, 1999

[4] M. A. Peabody, "Methods for pronunciation assessment in computer aided language learning," Doctoral dissertation, *Massachusetts Institute of Technology*, 2011

[5] C. Koniaris, " motivated speech recognition and mispronunciation detection," Doctoral dissertation, *KTH-Royal Institute of Technology Stockholm*, 2012.

[6] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech & Language*, 21(1), 219-230, 2007

[7] F. de Wet, C. Van der Walt and T. R. Niesler, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Communication*, 51(10), 864-874, 2009