Russian, Japanese, Arabic, Korean and Mandarin Chinese. The Pilot earpiece is connected via Bluetooth to the Pilot speech translation app, which uses speech recognition, machine translation and machine learning and speech synthesis technology. Simultaneously, the user will hear the translated version of the speech on the second earpiece. Moreover, it is not necessary that conversation would be taking place between two people; only the users can join in and discuss as a group. As if now the user may experience a few second lag interpolated the speech and translation, which Waverly Labs pursue to reduce. The Pilot earpiece will be available from September but can be pre-ordered now for $249. The earpieces can also be used for streaming music, answering voice calls, and getting audio notifications.

Link: https://www.indiegogo.com/projects/meet-the-pilot-smart-earpiece-language-translator-headphones-travel#/

4 Datasets in NLP and state-of-the-art models

The objective of this section is to present the various datasets used in NLP and some state-of-the-art models in NLP.

4.1 Datasets in NLP

Corpus is a collection of linguistic data, either compiled from written texts or transcribed from recorded speech. Corpora are intended primarily for testing linguistic hypotheses - e.g., to determine how a certain sound, word, or syntactic construction is used across a culture or language. There are various types of corpus: In an annotated corpus, the implicit information in the plain text has been made explicit by specific annotations. Un-annotated corpus contains raw state of plain text. Different languages can be compared using a reference corpus. Monitor corpora are non-finite collections of texts which are mostly used in lexicography. Multilingual corpus refers to a type of corpus that contains small collections of monolingual corpora based on the same sampling procedure and categories for different languages. Parallel corpus contains texts in one language and their translations into other languages which are aligned sentence phrase by phrase. Reference corpus contains text of spoken (formal and informal) and written (formal and informal) language which represents various social and situational contexts. Speech corpus contains recorded speech and transcriptions of recording and the time each word occurred in the recorded speech. There are various datasets available for natural language processing; some of these are listed below for different use cases:

1. Sentiment Analysis: Sentiment analysis is a rapidly expanding field of natural language processing (NLP) used in a variety of fields such as politics, business etc. Majorly used datasets for sentiment analysis are:

a) Stanford Sentiment Treebank (SST): Socher et al. introduced SST containing sentiment labels for 215,154 phrases in parse trees for 11,855 sentences from movie reviews posing novel sentiment compositional difficulties [127].
b) Sentiment140: It contains 1.6 million tweets annotated with negative, neutral and positive labels.