Methods for Spoken Language Identification

Julien Boussard, Andrew Deveau, Justin Pyron

{julienb, adeveau, pyron}@stanford.edu

Overview

For our project, we explored several machine learning techniques for classifying spoken language. In particular, we constructed algorithms which utilize various features derived from English and Mandarin Chinese phone call audio to predict the language to which the phone call belongs. We investigated multiple feature sets and modeling approaches, and found that Gaussian Mixture Models and Feed-Forward Neural Networks, combined with shifted delta cepstra (SDC) features, achieved the best performance.

Dataset

We used the OGI Multilanguage Corpus as our dataset [1]. This dataset consists of phone calls in ten languages, with durations ranging from a few seconds to approximately one minute. We used calls in English and Mandarin of at least three seconds in length, of which 60% percent were assigned to the training set and 20% were assigned to each of the validation and test sets.

Features

The core set of features we utilized to make predictions was Mel Frequency Cepstral Coefficients (MFCCs), which are perhaps the most widely utilized feature in digital signal processing for capturing speech information. Each sound wave is composed of a set of sinusoids of different frequencies, and, roughly speaking, the MFCCs capture information about the energy in each frequency, adjusted to account for the way humans perceive sound. Each phone call consists of a time series of MFCC vectors, each of which is computed from one of the 25-millisecond frames into which each speech segment is broken. **See figure 1.**

To capture information about how the speech signal evolves through time and relates to nearby speech utterances, we additionally included delta cepstra and shifted delta cepstra as features [2]. These features are constructed by taking differences between MFCC vectors. **See figure 2.**

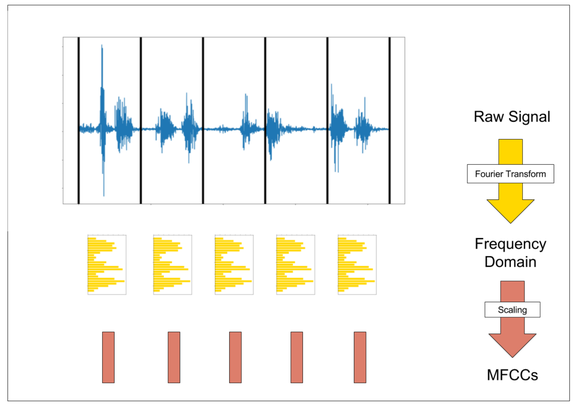

Figure 1: Feature construction

Gaussian Mixture Model (GMM)

Each language contains a fixed number of distinct phonetic expressions. Under this modeling approach, it is assumed that each of these distinct expressions has associated with it a Gaussian distribution, and that each utterance is drawn from one of them. For any given utterance, the phonetic expression to which it belongs is an unobserved latent variable, and the EM algorithm must be employed in order to obtain the probability of that utterance occurring [3]. This probability is given by

$$\mathbf{P}(x^{(i)}|l) = \sum_p \mathbf{P}(x^{(i)}|p)\mathbf{P}(p|l)$$

To make call-level predictions, a mixture model is first fitted for each language separately. For each utterance in a phone call (an MFCC vector) the probability of that utterance occurring in each language is computed via the GMM. After modeling $p(x^{(i)}|l)$ in this way, we tried two approaches for aggregating frame-level predictions into call-level predictions. We first used a majority vote, where a call was classified to the language with higher probability in a majority of the frames. We then tried making the assumption, akin to that of Naive Bayes, that the features for each frame in a call are generated by independently sampling from the learned Gaussian mixture

$$\mathbf{P}(x^{(1)}, x^{(2)}, \ldots, x^{(m)}|l) = \prod_{i=1}^{m} \mathbf{P}(x^{(i)}|l)$$

We then classified via

$$\arg\max_{l \in \{E,M\}} \prod_{i=1}^{m} \mathbf{P}(x^{(i)}|l)$$

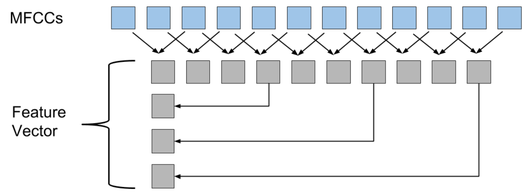

Figure 2: Shifted Delta Cepstra

Feed-Forward Neural Network (FFNN)

As an alternative approach to capturing the time dynamics of speech, we trained a feed-forward neural net on a feature set that additionally contained shifted delta cepstra. Call-level predictions were formed with the same majority vote method. Such a modeling framework yielded comparable results to a CNN trained only on static MFCC features. A variety of architectures were tested, and shallower architectures with a large number of neurons generally achieved the best performance.

Convolutional Neural Network (CNN)

We also trained a convolutional neural network to make predictions directly on the call level. Training examples were two dimensional arrays consisting of all the MFCCs for a particular call. This model is also equipped to capture time dynamics of speech by linearly combining MFCCs from different frames. The architecture and L^2 regularization parameters were chosen via a grid search, though the validation accuracy proved fairly insensitive to the choice of these parameters.

Results

We performed binary prediction on English and Mandarin phone calls. The results obtained from various modeling approached can be found in **table 1.**

Model	Train Accuracy (1303 calls)	Test Accuracy (449 calls)
GMM	**95.9%**	**86.6%**
FFNN	**73.5%**	**70.8%**
CNN	**98.7%**	**68.0%**

Table 1: Model Performance

Discussion

Our preliminary testing involved making predictions utilizing only static MFCC feature vectors, combined with simples statistics of their distribution in each call (e.g. mean, variance). The resulting models performed poorly, and often performed worse than random guessing. The failure of these models highlights the importance of granular time dynamics. To a large extent, a language is not distinguished by the presence of certain sound waves, but rather by the patterns they form and the sequence in which they are produced. By utilizing SDC features along with GMM and neural network models, we were able to effectively capture this crucial information, leading to improved predictive power.

Future

There exists a rich set of literature describing techniques for extracting information from a signal, but their successful implementation requires an advanced and thorough knowledge of signal processing. Given more time, we would like to explore this area in greater depth, which would allow for enhanced data cleaning and normalization, as well the inclusion of additional features. Helpful features would capture information such as the rhythm and modulation of speech. It would also be interesting to train a CNN on a larger data set and to understand how its learned representation compares to the representation used by the GMM.

References

[1] Y.K Muthusamy R.A Cole.
"The OGI Multilanguage Telephone Speech Corpus".
Proceedings International Conference on Spoken Language Identification, vol. 2 pp. 895-899 Oct. 1992.

[2] L. Wang B. Yin E. Ambikairajah, H. Li and V. Sethu.
"Language Identification: A Tutorial".
IEEE Circuits and Systems Magazine, vol. 11, no. 2, pp. 82-108, 2011.

[3] Reynolds Torres-Carrasquillo and Deller.
"Language Identification Using Gaussian Mixture Model Tokenization".
IEEE International Conference on Acoustics Speech and Signal Processing, pp. 757-760, 2002.

[4] K. M. Ting Z. Fu, G. Lu and D. Zhang.
"A Survey of Audio-Based Music Classification and Annotation".
IEEE Transactions on Multimedia, vol. 13, no. 2, pp. 303-319, 2011.