Language Identification using Spectro-Temporal Patch features

Kamal Sahni[1], *Pranay Dighe*[2], *Rita Singh*[3], *Bhiksha Raj*[3]

[1]Department of Electrical Engineering, Indian Institute of Technology Kanpur, India
[2]Department of Computer Science & Engineering, Indian Institute of Technology Kanpur, India
[3]Laugauge Technologies Institute, Carnegie Mellon University, USA

skamal@iitk.ac.in, pranayd@iitk.ac.in, rsingh@cs.cmu.edu, bhikhsa@cs.cmu.edu

Abstract

We present a novel approach for automatic Language Identification (LID) using spectro-temporal patch features. Our approach is based on the premise that speech and spoken phenomena are characterized by typical visible patterns in time-frequency representations of the signal, and that the manner of occurrence of these patterns is language specific. To model this, we derive a randomly selected library of spectro-temporal patterns from spoken examples from a language, and derive features from the correlations of this library to spectrograms derived from the speech signal. Under our hypothesis, the relative frequency of correlation peaks must be different for different languages. We model this by learning a discriminative classifier based on these features to detect the presence of the language in a recording. The proposed approach has been tested on two different datasets: the VoxForge multilingual speech data and CallFriend corpus available from the Linguistic Data Consortium (LDC).

Index Terms: Language identification, Spectro-temporal patches, Discriminative classification.

1. Introduction

Language Identification is a problem of identifying the language in a spoken utterance. It has many applications, such as for categorization of audio material, front-ends for multilingual speech recognition systems, automatic customer routing in call centers of different companies etc.

The most successful approaches to automatic language identification thus far have explicitly utilized the phonotactic structure of the spoken language. For instance, phone-recognition based approaches [1], [2] compute the score for any language through a phoneme recognizer that is guided by an N-gram language model for the phonemes in the language. Parallel phone recognition based techniques [3], [4] go a step further and simultaneously recognize the speech using phoneme recognizers for multiple languages, and utilize the ensemble of outputs to identify the language. LVCSR based systems perform entire large vocabulary recognition [5]. In each case, the identification of the language is based on matching entire phoneme-level spectral patterns in the incoming speech to known patterns for the language to identify it.

Regardless of the success of phoneme-based methods, it is generally also acknowledged that information about the identity of the language is also present in the spectro-temporal patterns in the signals, evidenced partially by the fact that humans can often identify a language even when they do not have a working knowledge of the language. Consequently, a large number of purely *acoustics* based methods for language identification have also been proposed in the literature. GMM-based methods [7] only model the distribution of individual spectra of recordings from the signal. However, even acoustically, it is generally understood that the information actually lies in longer-range patterns. Consequently Pedro et al.[8] have modeled the sequence of Gaussian indices obtained for individual frames of a recording from a GMM. Ma et al. [9] use automatically defined acoustic segment units to model the distinction between languages. In all of this too, the patterns that are modeled are still *spectrally complete* – they only vary in their *temporal* extent.

In this paper we propose to exploit an entirely different scale of feature. We hypothesize that the information about the underlying message in a speech signal also lies in *local* spectro-temporal patterns in the signal. A significant aspect of the distinction between different languages lies in the nature and manner of occurrence of these patterns. By appropriately characterizing the patterns and the rate and manner in which they occur, we can therefore expect to identify the language being spoken. We note that a similar hypothesis has previously also been explored by Ezzat et al.[6] for word spotting.

The above hypothesis would argue that in order to identify the spoken language properly, we must therefore know about the spectro-temporal patterns in all candidate languages that may have been spoken. However, motivated by the fact that humans can often *detect* a segment of familiar sounding language even in the midst of a stream of otherwise unrecognizable gibberish, we pose the problem differently: as one of merely determining if the patterns typical for a given language occur or not. Thus our solution is more appropriately called language *detection* rather than identification.

To learn the spectro-temporal patterns and their occurrence patterns automatically, we use an approach similar to that in [6]. We derive a large number of randomly chosen spectro-temporal patterns from examples of the language. We characterize the rate of occurrence of each of these patterns through their correlation to spectro-temporal representations of the signal. Finally, a discriminative classifier employs these characterizations as features for classification.

Preliminary results on two different databases indicate that the proposed approach is able to perform very accurately on detecting a target language even in snippets of speech that are 10 seconds or shorter in length.

We have performed a comparison of our approach with the work done by Campbell et al. [11], in which SDC (shifted delta cepstral coefficients) have been used as feature vectors. The comparison shows that the proposed approach is competitive. Notably, the features we use are very dissimilar to those in Campbell et al. Presumably, combining the two could result in even better performance.

The rest of the paper is arranged as follows. In Section 2 we describe the overall rationale behind the use of spectro-temporal patterns for language identification. In Section 3 we outline our

mechanism for learning spectro-temporal patch dictionaries. In Section 4 we describe how we use these to derive features from the speech data. Section 5 describes our classification strategy, Section 6 presents our experiments and in Section 7 we give our conclusions.

2. Rationale behind spectro-temporal patches for language identification

2.1 The information in speech is represented in its spectro-temporal patterns

It is well known that the identity of a speech sound is evident from the spectro-temporal patterns in spectrographic representations. In fact, many early speech recognition systems attempted to utilize this characteristic by explicitly attempting to "read" the spectrogram. Later research veered away from this approach to frame-based statistical characterizations that only explicitly represented the spectral characterizations, leaving the representation of temporal characteristics to an underlying Markov chain in a hidden Markov model.

Although several researchers have attempted to revisit explicit spectro-temporal characterizations, these approaches have largely not resulted in significant improvement over the HMM approach, primarily because they remained tied to a state-based characterization [SSMs] or to length restrictions in patterns [STMs] and also generally ignored the fact that the patterns in the speech spectrogram include both frequency-localized long-term patterns that extend over several tens of milliseconds and short-term patterns that are local not only in frequency, but also in time.

Yet it remains true that the typical local patterns such as formant trajectories etc. in spectrograms remain visible even in high levels of noise, even when the individual spectral vectors in the signal are corrupt beyond recognition. It also remains true that these patterns characterize nearly the totality of the information in the speech signal, including the identity of the underlying phonemes, the speakers, and the language being spoken.

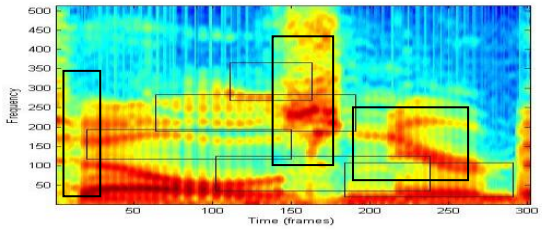

Figure 1: *Sample patches from a spectrogram*

2.2 Characterizing speech through local spectro-temporal patterns

In this paper we therefore revisit the use of explicit characterizations of spectro-temporal patterns in the speech signal to perform pattern classification tasks on speech, specifically that of identifying language.

Our approach is based on the following observations:

a) The identity of the sound in any speech recording is encoded in the spectro-temporal patterns that occur in it.
b) These patterns are local in the time-frequency plane.
c) The identity of a language is encoded in the pattern of occurrence of these spectro-temporal patterns.

We will, however not attempt to identify the specific patterns that are most useful. Instead, we will hypothesize a large number of them and determine their relevance to the task at hand in a data driven manner.

3. Spectro-temporal patch dictionary

As mentioned above, we do not attempt to identify the most relevant spectro-temporal patterns explicitly. Instead, we hypothesize a large number of candidate patterns, all of which are likely to carry relevant information.

To represent the spectro-temporal patterns in any language, we create a *patch dictionary*, consisting of randomly chosen rectangular spectro-temporal patches of random sizes, from the spectrogram of a relatively small amount of *exemplar* training data from the language. These patches are extracted from random locations in time and frequency in the spectrograms of the exemplar data. The height and width of each patch (representing its span along the frequency and time axes) are chosen randomly from a spectral range F_{range} and a temporal range T_{range} respectively. Finally, all patches with a total energy below a threshold are discarded, to ensure that all patches that are extracted have some energy in them; otherwise we might end up with a lot of empty patches that carry little or no acoustic information. The remaining patches are stored in the dictionary *along* with the frequency location from where they were derived. Figure 1 illustrates patch extraction from a spectrographic representation of an audio file. Given a sufficiently large number of patches, several of them will capture many types of typical spectro-temporal phenomena, such as formant ridges/sweeps, harmonic lines, noise patterns, etc., some of which will be characteristic of the language.

4. Patch based feature extraction

The library of spectro-temporal patches can now be used to derive features from any spectrogram. We employ each of the patches as a *matched filter* on the spectrogram. We correlate the patch with the entire strip of the spectrogram that represents the same frequency range as the patch. Peaks in the correlation indicate matches, indicating occurrences of the patch. Enumerating these gives us an indication of the rate of occurrence of the patch within the spectrogram. This is illustrated in Figure 2. Let $P_i(f,t)$ represent the i^{th} patch in our library. Let there be M patches in our dictionary. Our extracted patch dictionary can hence be represented as $\boldsymbol{P} = \{P_i(f,t) : i = 1...M\}$. We will now use this dictionary to compute the feature vector for any speech recording.

Let $S(f,t)$ represents spectrogram of a signal s. Let, T be the total length of the spectrogram. Let W_m and H_m be the width and height of the m^{th} patch in the dictionary. Let F_m be the frequency location from which it was drawn.

In order to compute a match between $P_m(f,t)$ and the signal s, we compute the cross correlation between $P_m(f,t)$ and the portion of the spectrogram $S(f,t)$ that covers the same frequency range as $P_m(f,t)$, i.e. the *sub*-spectrogram $S_m(f,t) = S(f,t) \mid F_m <= f <= F_m + W_m$. In principle, this could be computed very fast using a 2-D fast-Fourier transform, however such a computation would ignore local variations in signal level differences and results in poor characterization of the signal. Instead, we use a *normalized*

2-D cross-correlation to characterize the match. Moreover, in order to account for the fact that the precise location of spectrographic pattern along the frequency axis may vary from speaker to speaker, depending on their gender, the length of their vocal tract etc., it is not sufficient to merely compute the correlation in the frequency range F_m instead we consider an *extended* frequency range(F_m–Δf/2, F_m+Δf/2), and use the peak correlation within this range as the overall normalized cross correlation at each instant.

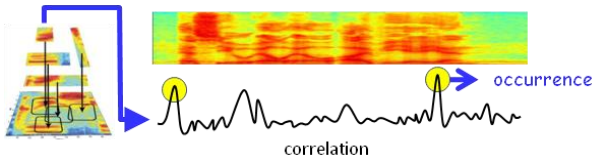

Figure 2:*Patches are correlated against the strip of the spectrogram from the same frequency range (the black rectangle). Peaks in the correlation indicate occurrences. Part of the image is taken from [6].*

Thus, for the patch $P_m(f,t)$, we obtain the normalized cross-correlation at any time t_n as:

$$C_m(t) = \max_{f \in (F_m - \Delta f, F_m + \Delta f)} R_m(f,t)$$

Where,

$$R_m(F,T) = \frac{\sum_{f,t}[\,(S_m(f-F,t-T) - \overline{S_m}(F,T)).(P_m(f,t) - \overline{P_m})\,]}{\sqrt{\sum_{f,t}(S_m(f-F,t-T) - \overline{S_m}(F,T))^2 \cdot \sum_{f,t}(P_m(f,t) - \overline{P_m})^2}}$$

and where $\bar{S}(F,T)$ is the mean of the spectrographic region given by $S(f,t) \mid F <= f <= F+W_m$ and $\overline{P_m}$is the mean of the patch.

This results in a sequence of $C_m(t)$ values representing the normalized cross correlation function between $P_m(t,f)$ and $S(t,f)$ and is a series of cross-correlation values. From this series 3 numbers are computed: the mean, the variance, and the number of times it exceeds a threshold. The threshold used in this experiment is 0.6 per unit time, representing the rate of detection of the patch. These three values are derived for *every* patch in the dictionary. Thus, for a dictionary with *M* patches, we derive a *3M* dimensional feature vector for every speech recording.

5. Discriminative classification via SVM

As mentioned in the introduction, we actually perform a binary language detection task, rather than multi-class language identification; however the procedure is easily extended to multi-class classification as well. The 3M dimensional patch-based features are now used in a discriminative classifier. We obtain a collection of within-language and out-of-language recordings as positive and negative exemplars, derive feature vectors for all of these, and train a support-vector machine [10] from the collection. The theory of support vector machines is well known and need not be repeated here. Thereafter, for each test utterance that must be classified as being from the target language or not, we derive a 3*M* dimensional feature vector as described and classify it using the SVM. In our experiment we used a SVM with a linear kernel.

6. Experimental Results

We evaluated our proposed technique on two corpora: The CallFriend corpus available from the Linguistic Data Consortium (LDC) and the VoxForge multilingual dataset obtained from voxforge.org. The LDC data are quite noisy data and recorded over a telephone, whereas VoxForge data are comparatively clean in terms of background noise etc.

6.1 System parameters

Spectrograms were computed as log-magnitude short-time Fourier transforms with 25ms analysis windows and 6.25ms frame-shifts. For our experiments the temporal range used for the width of the patches was [0.1, 0.6] sec. The spectral range used was F_{max}[0.1, 0.4], where F_{max} is the highest frequency in the spectrogram. In all experiments, the patch dictionary was composed from positive speech examples from approximately 15 minutes of data from the language to ensure good variability in terms of sounds present in the language.

To train the SVM an additional 80min each of within-language and out-of-language data were used. For all tests, 30 minutes each of within-language and out-of-language data were used. In all experiments, the data were chopped into segments of no more than 10 seconds in length. It is therefore worth noting that all reported results are from segments of speech that are *no more than 10 seconds* at a time, and are often much shorter.

6.2 Effect of dictionary size

Since patch based features are crucial for our methodology, in a preliminary experiment we analyzed the effect of dictionary size on this language identification. Here we employed the LDC corpus and trained the classifier to detect English. All other languages were treated as negative instances. Results were obtained for different numbers of patches in dictionary. The number of patches extracted was varied from M = {300, 600, 1000, 1500, 2000, 2500, 3000, 4000, and 5000}. Figure 3 is a plot of EER (equal error rate) as a function of the number of patches. Clearly, as the dictionary size increases, error decreases and the performance of the classifier increases. In subsequent experiments we used a dictionary size of 4000 patches.

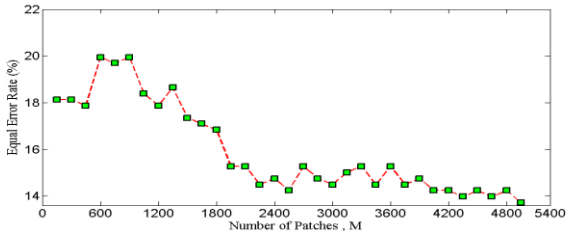

Figure 3:*Equal Error Rate (%) vs. Number of Patches*

6.3 LID results on CallFriend - LDC data

In next experiment we used four languages from CallFriend-LDC corpus: English, German, Hindi and Farsi. We made 4 SVM based binary classifiers, designed for detection of each of the four languages.

For each language, patch-composition, training and test data were set up as described in Section 6.1. For each language, the negative data were assumed to comprise the remaining 3 languages. Data used to learn the patches were not used to train the classifiers.

Detection Error Tradeoff (DET) is shown in figure 4, for patch based classifiers, for each of four languages. EER (equal error rate) for each of the 4 languages is tabulated in Table 1.

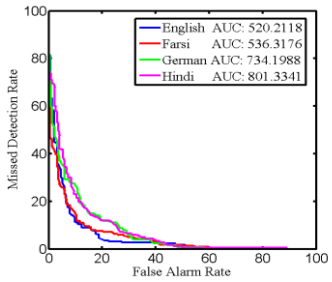

Figure 4: *DET plots for 4 different languages: English, Farsi, German and Hindi, taken from CallFriend Corpus.*

Language	EER
English	10.77%
Farsi	11.32 %
German	15.05 %
Hindi	15.16 %

Table 1: *EER performance of the systems on CallFriend Corpus*

The average EER over all the languages from the patch based approach is 13.1%. Campbell et al.[11] report an EER of 6.1% using shifted delta cepstral coefficients (SDCs) over 12 languages from CallFriend, treating it however as a multi-class language *ID* problem, and using gender-specific models. Unlike Campbell et al., we perform *detection*, a more difficult task. Also we don't require separate models for males and females.

6.4 LID for different users speaking a particular language

The VoxForge multilingual dataset includes speech examples for a number of users. We have taken English and Russian language data for this experiment. For this data we built a detector for English – we note however that since there are only two languages in the corpus, this reduces to a relatively simple binary classification task; nevertheless classification task: English vs. Russian does provide some additional information. Additionally, the test provides a measure of performance on broadband data, and is particularly useful since one of our test speakers had a distinctly non-native accent, thus giving us an idea of the consistency of spectral-pattern based classification across accents.

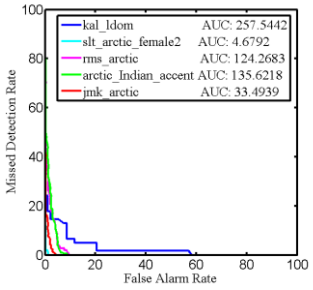

Figure5: *DET plots for 5 different English speakers.*

Speaker	EER
kal_ldom	8.6%
jmk_arctic	2.6%
rms_arctic	5.1%
arctic_Indian_accent	4.9%
slt_arctic_female2	1.3 %

Table 2: *EER performance of the systems with different speakers, on the VoxForge multilingual dataset.*

The training data included 2 males and 1 female speaker with typical American accent. The test data included 3 male speakers (kal_ldom; rms_arctic; jmk_arctic) and one female (slt_arctic_female2) speaker with typical American accent, and one male (arctic_Indian_accent) with Indian accent. Negative data, both for training and test had a number of speakers of both genders from the Russian language exemplars.

The DET for the performance obtained with the patch-based system is shown in Figure 5. EER performance of the system over different speakers is tabulated in Table 2.

7. Conclusions and Future work

We present a novel language identification technique based on characterization of the relative frequency of occurrence of spectro-temporal patterns. Experiments show that this technique can provide competitive performance in comparision to the approach proposed by Campbell et al. [11]. Since both the approaches use entirely different kind of feature sets, the two could potentially be combined to get better results than either.

We aim to improve the performance of our technique over the reported results. We have not investigated the optimization of spectro-temporal representations. Since a large number of patches have been used in this experiment, the features are expected to be redundant. Dimensionality reduction techniques may be employed to get better results.

A particular feature of our approach is that it can work with very short segments of audio. Thus is a particularly promising tool to detect code switching. These, and the extension to multi-class classification, are objectives of future work.

8. REFERENCES

[1] T. J. Hazen and V. W. Zue, "Automatic language identification using a segment-based approach," in Proc. Eurospeech '93, vol. 2, Sept. 1993, pp. 1303-1306.

[2] R. C. F. Tucker, M. J. Carey, and E. S. Paris, "Automatic language identification using sub-words models," in Proc. ICASSP '94, vol. 1, Apr. 1994, pp. 301-304.

[3] L. F. Lame1 and J.-L. Gauvain, "Identifying non-linguistic speech features," in Proc. Eurospeech '93, vol.1, Sept. 1993, pp. 23-30.

[4] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in Proc. ICASSP '94, vol. 1, Apr. 1994, pp. 305-308.

[5] T. Schultz, I. Rogina and A. Waibel, "LVCSR-BASED LANGUAGE IDENTIFICATION" in Proc. ICASSP '96, vol. 2, May 1996, pp. 781-784.

[6] T. Ezzat and T. Poggio, "Discriminative word-spotting using ordered spectro-temporal patch features," in Proc. SAPA, Brisbane, 2008.

[7] Marc A. Zissman, "Comparison of :Four Approaches to Automatic Language Identification of Telephone Speech", IEEE, 1996.

[8]Pedro A. Torres-Carrasquillo, Douglas A. Reynolds and J.R. Deller, Jr. "Language Identification using Gaussian Mixture Model Tokenization", IEEE, 2002.

[9] L. Haizhou; M. Bin and L Chin-Hui "A Vector Space Modeling Approach to Spoken Language Identification," in Proc. Audio, Speech, and Language Processing, IEEE, Vol. 15, Issue 1, pp. 271 - 284

[10] Nello Cristianini and Bernhard Sch¨olkopf, "Support Vector Machines and Kernel Methods: the New Generation of Learning Machines," AI Mag., vol. 23, no. 3, pp. 31–41, 2002.

[11] W. M. Campbell, P. A. Torres-Carrasquillo, and D. A. Reynolds, "Language recognition with support vector machines," in Proc. of Odyssey 04, 2004, pp. 285–288.