The Unsolved Problem of Language Identification: A GMM-based Approach

Maggie Mi
Lancaster University

Abstract. Language identification (LID) systems attempt to identify a language from a series of randomly spoken utterances (Das & Roy, 2019), and this provides the foundation of many natural language processing (NLP) applications, such as multimedia mining, spoken-document retrieval, as well as multilingual spoken dialogue systems (Navratil, 2006). However, presently, the LID task is still an unsolved problem, often with increasing equal error rate (EER) as the duration and quality of the dataset decreases (Ambikairajah et al., 2011). The HMM-GMM (Hidden Markov Model-Gassian Mixture Model) approach taken in this paper involves building an acoustic model that uses probabilistic representations of speech datasets across 10 languages (Dutch, Russian, Italian, Portuguese, German, English, French, Turkish, and Greek). Through the exploration of the cross-linguistic features present in language families and the effect of the experimental parameters on the performance of the system, i.e., the length of the data recording, areas of weaknesses and corresponding means of improvement are therefore revealed.

Keywords: Spoken language identification; speech processing; HMM-GMM; computational linguistics; natural language processing (NLP)

1 Introduction

1.1 Background

Automatic Language Identification (LID) is defined as, 'the identification a language from random[ly] spoken utterances' (Das & Roy, 2019, p.81). Although this technology is also available for written text, for example, Google Translate's 'Detect language' function, the focus of this project is from a speech technology standpoint. The implementation of LID systems can be seen throughout many different areas. For instance, technological conglomerates require such a system to aid the categorisation of user data. Completing this manually will not only be labour intensive but also costly in terms of human resources and time efficiency (Baldwin & Lui, 2010). Telephonically, LID systems have been often used for emergency services and call centres, including parts of the tourism industry (Muthusamy, 1993). The shared purpose of these systems is the ability to route the caller to the right interpreter, or agents speaking the same language, to help the caller with their needs. Thus, the emphasis on accuracy is particularly important, especially, in life-and-death situations dealt with by emergency services.

What makes this field particular intriguing, is the attempt at capturing the characteristic and representative features of natural language, through speech data, which is highly idiosyncratic and inconsistent, with the quality of the data directly affecting the performance of such systems (Xu, Ding, &

Watanabe, 2019). Speech, being less tangible than textual data, is temporal. It is also liable to an individual's articulatory differences. For example, younger speakers tend to have high F1 and F1 values, similarly, in general, the vocal tract length of males is greater than females, thus, giving rise to inevitable variety (Kumar et al., 2011; Rastatter et al., 1997).

Taking the International Phonetics Alphabet as a framework, it is proposed that, there are in total 107 phones deemed producible by the human anatomy. Combined with 4 prosodic markers of intonation and 52 diacritics marks, this means, there are a finite number of possible combinations of sounds, and these combinations often overlap in many languages (Association & Staff, 1999). An example would be 'bùkě' (不可), meaning 'must not; do not', as opposed to 'book' in English. Being typologically different languages, Chinese, a Sino-Tibetan language, has a system of intricate tones, which are absent from the (Proto-)Indo-European languages, such as English. However, despite intonation differences, there also exists phonemically similar utterances across languages, regardless of the typological language family distinction. Even in English, phrases such as, 'Let's recognize speech' and 'Let's wreck a nice beach' shares a similar combination of phonemes and thus, at times, even humans can perceptively misinterpret the two. Therefore, the project aims to explore the following areas: the pattern of recognition performance between typologically similar languages, weaknesses of such a system and means to improve these shortcomings, and finally, whether experimental methodology and parameters, such as duration of the data would affect classification results. The approach taken to explore these research questions involves building a GMM-HMM LID system. Compared to rule-based systems, which aims to break down a language into different parameters and corresponding rules and assumptions, GMM-HMMs are conceivably more advantageous in terms of capturing an overall representation of a language (Ives, 1986; Zissman, 1993).

In the subsequent sections, the system design will be provided, setting out the inner workings of the system (Section 4). This is followed by descriptions of the systems' performance in Section 5. Sections 6 then aims to analyse the results and discusses ways to improve the shortcomings of the system. The goal of this project is to obtain an in-depth understanding of this statistical approach towards LID, and to overcome the issue of misclassification, as much as possible.

2 Literature Review

Initial studies of LID systems date back to the 1970s. Since then, a various range of techniques has been developed. In this section, a brief history of previous methodologies used in building LID systems will be analysed with reasons for choosing GMM-HMM modelling as the methodology of this project.

In order to narrow down the scope of the focus placed on the literature; certain factors were kept in mind when completing this review. These include:

(A) Number and similarity of languages (i.e., language families): The fewer the languages, the easier the recognition task; similar languages are harder for identification.

(B) Data used for the study: whether there is an overlap in speakers present in test and training data; gender-inclusiveness in the data; quality of data (i.e., recording conditions) …etc all these features will directly affect the meaningful conclusions, that can be drawn from a study.

(C) Methodology: HMMs-based approaches, clustering, neural networks…etc

2.1 Static Classification Methods

Leonard and Doddington (1974) were believed to be the earliest scholars of this field. Their approach to the problem involved extracting filter bank feature vectors and spotting regions of change and stability. Filter bank feature vectors is a type of quantified representation of an acoustic signal. A filter bank typically contains sets of bandpass filters that permit certain frequencies of a specified range but rejects frequencies outside that range. Each bandpass filter can be thought of as being able to 'select' or 'distinguish' a particular band of frequencies from the input signal. The feature vector produced by the filter bank is a vector, or conceptually, a string of numbers, representing 'the amount of energy in each frequency band' (King, 2017). Thus, by inspecting the filter bank feature vectors, such regions of change and stabilisation were taken to be indicative of a specific language. Accordingly, these regions were then used as templatic patterns for comparison with the test data.

Due to the classified nature of this research the languages used were not disclosed, nor was background information of the 100 speakers. Thus, it is difficult to come to evaluate the 70% correct classification across the five languages without additional information such as typological, or phonemic similarities in these languages, or metadata of the participants.

2.2 Importance/Presence of Linguistic Units

Later studies completed by Leonard and Doddington cemented the idea that linguistics units are most likely to differentiate languages. Due to the varied occurrences of linguistic units in different languages, it was hypothesised that using the distributions of frequency of such units, different languages can be distinguished. As sounds can be represented in terms of their acoustic representation, thus, extracting features from the speech signal retains the unique features of the language (Leonard & Doddington, 1974; 1978)[42].

Cimarusti and Ives (1982) found that discriminating features are not necessarily linguistic features. Their approach disregarded the idea of linguistic units, such as syllables or phones, and instead, they implemented pattern analysis techniques to acoustic features extracted from the speech signal. The overall result of the system was 84% performance accuracy, with scores ranging from 76.8% (American English) to 93.4% (Korean). From these results, it is plausible to conclude that, acoustic features alone can be used in LID. Although, this would need to be further tested as the dataset has a small sample size of 5 speakers per language. This also leads to the speculation as to whether the system is "speaker-independent", meaning the system could be picking out characteristics of the speaker as opposed to the language.

In a subsequent study conducted by Ives (1986), both prosodic feature vectors and formants values used as a basis for a LID system were explored. It was found that formant frequencies are more useful in distinguishing languages. Thus, classification was based on thresholds and densities of quantified representation of each language. LID was then based on sets of 'production rules', defined by variance, the value of F0 and variance of F2, amongst other parameters. Classification results of this system had an overall accuracy of 92%, and individual classification results ranging from 84% (Russian) to 99% (Vietnamese). Despite the promising results, the study lacked information on the amount of data and test/training data assignments. Moreover, the possibility of an overlap between training and test sets were not eliminated. Given the approach taken in building this system is based on formants, it is critical for the system to be exposed to speech produced by female speakers. This is because, acoustically, female speakers

[42] This foundational assumption has fuelled many other works in the field, such as Cimarusti and Ives (1982) and Foil (1986).

tend to have higher F0 values than males, thus, it would be interesting to see how the system would perform when exposed to female speakers (Muthusamy, 1993).

2.3 (GMM-)HHM-based Studies

A prominent study following Leonard and Doddington (1974) is completed by House, Neuberg, and Wohlford (1975). In this leading study, House and colleagues proved the possibility of using sequences of phonetic categories (stops, fricatives, vowels, silences) to tackle LID. Working under the assumption, that these categories can be demonstrated as a Markov process, and that the model's parameters could be projected from necessary training data, phonetic transcriptions were used for the following eight languages: Urdu, American English, Greek, Chinese, Japanese, Swahili, Korean and Russian. Phonetic category labels from these phonetic transcripts were then used to train statistical models. The underlying concept is given a sequence of category symbols U, and language L, P(U|L) is calculated for each language. 'U is said to represent the language L, for which this probability is a maximum' (House et al., 1975; Muthusamy, 1993, p.14).

The study of House et al. (1975) is an important underpinning foundational work for many future designs of the LID and speech recognition systems. However, in their actual experiment, speech recordings were not used, only phonetic transcriptions of texts from each of the eight languages. This is a major flaw in methodology design, as 'absolute', or 'perfect' acoustic segmentation of real speech data is very much unfeasible (Muthusamy, 1993). The continuous sinusoidal nature of speech signals' frequencies further proves this. Thus, this experiment is not representative of the performance and effectiveness of the system on actual speech data.

Li and Edwards further advanced the grounding HMM techniques by House et al. (1975) and used them in real speech data (Li & Edwards, 1980). They used a segmentation framework that contains six categories: syllabic nuclei, non-vowel sonorants, vocal murmur, voiced frication, voiceless frication and silence, or low energy segments. The segmental models characterised the likely, segmental sequences in the language, whereas the syllable model was split into two types: inter-syllable-nuclei sequences and intra-syllable-nucleus segment sequences. The former can be thought of as depicting the likelihoods of possible consonant clusters, and the latter, as the likelihood of a certain internal arrangement of a syllable.

The training data consisted of 200 minutes from 20 speakers of five languages, three Indo-European and two Asian, reading aloud. Due to the nature of the Asian languages being tonal and monosyllabic, the utterances follow straightforward consonant-vowel (CV) or consonant-vowel-consonant (CVC) formations. In contrast, the rest of the dataset is far more complex, with compounded consonant clusters and longer word lengths. A maximum of approximately 80% correct identification was gained using the inter-syllable model. Analysing the mistakes made by the system revealed that, this approach excellently differentiated the Asian languages from Indo-European languages (Li & Edwards, 1980).

As an extension on the work of House et al. (1975), this study supports the hypothesis that, 'broad phonetic category sequences do possess language discriminatory information' (Muthusamy, 1993, p.16). However, this study contains several caveats, such as the limited information on the languages used, quality of recording conditions…etc are all unknown factors. Moreover, the study lacks consideration of variability as expected in LID systems. Namely, the nature of reading speech and the absence of female speakers greatly limits the reproducibility of the results in real-world applications.

A final notable study that needs to be addressed is one completed by Nakagawa, Ueda, and

Seino (1992). In this study, four different approaches to LID are compared: vector quantisation (VQ), discrete HMM, continuous density HMM and a mixture Gaussian distribution model.

> 'A mixtured Gaussian distribution model is regarded as a special case of a continuous HMM with mixtured distributions, that is, the number of states corresponds to only one and the distribution is Gaussian.' (Nakagawa et al., 1992, p.1012).

Conceptually, this can be thought of as modelling the probability of a language occurring using normal distributions[43].

The dataset consisted of 750 recordings from four languages: English, Japanese, Mandarin Chinese, and Indonesian. Each language had 15 native speakers, producing 50 sentences e ach, with an average duration of the utterance at approximately, 3 seconds. Results from the continuous HMMs and GMM-HMM approaches yielded the best results, both at 81.1%, and were far more superior than the VQ (77.4%) and discrete HMMs approaches (47.6%). Despite this accomplished result, the generality of this study is once again limited by the lack of female speakers.

The studies reviewed above in 2.1-2.2, all primarily performed 'static classification'. HMMs, on the other hand, are more dynamic. They can 'model the sequential characteristics of speech production and have been used widely in speech recognition systems' (Zissman, 1993, p.399). As noted previously, languages differ from each other due to differences in phonemic inventories, as well as the realisation of similar phonemes in particular languages. For example, the fricative in German 'ich', has no Italian counterpart. Another example would be the /r/ in Spanish differing from its English counterpart (the former is realised as a trill, whereas the latter is a flap). Therefore, due to the complexity and intricacy of these small differences, it is more logical to use a comprehensive statistical modelling approach, to capture the characteristics of a language, rather than specified, preselected features.

Furthermore, in HMM-based LID systems, a language is identified by estimating the likelihood of each language occurring at 'contiguous, frames of the speech signal' (Radha, 2012, p.1101). Unlike static models, HMMs can be trained to represent different items, thus, this flexibility in defining the scope of a 'unit', means that HMMs are advantageous in building speaker-independent and text-independent systems. This, coupled with the temporal nature of speech data, and the ability to encode a language as a string of quantified representations, in the form of a spectral vector, makes GMM-HMM the preferred approach in building a LID system (Gales & Young, 2008).

3 Data

A common trend that is visible across the literature, is that the number of languages present in these systems usually does not exceed eight. There is a visible balance between building a system that deals with a high degree of complexity (i.e., multi-lingual compatibility) and one that has a high accuracy of classification. In the GMM-HMM system of Zissman (1993), one of the datasets used to train and test the system was the 20 languages CCITT database. Notably, the classification result of this system was the lowest, at 54%. On average, however, previous studies completed tend to have 2-8 languages. Therefore, an important aspect is to explore the shortcomings of these systems that aims to classify numerous languages. With this goal in mind, this project aims to sample as many languages as possible, eventually arriving at 10 languages.

[43] Please see Section 4 for a detailed account of how GMM-HMMs work.

3.1 Data Collection

Voxforge is a free open-source speech database, where people from all around the world voluntarily contribute speech data for the development of open-source speech recognition systems[44]. The repository offers speech datasets in the following languages: Albanian, Bulgarian, Catalan, Croatian, Dutch, English, French, German, Greek, Hebrew, Italian, Mandarin Chinese, Persian (Farsi), Portuguese, Russian, Spanish, Turkish, and Ukrainian. However, there is great variability in the datasets in terms of availability. For example, at the time of completion of this project, the Mandarin Chinese dataset is not available for download. On a similar note, within the datasets that are available for download, there exists a clear discrepancy between the magnitude of the datasets- Croatian, for instance, only contains two speakers, whereas English has 1234.

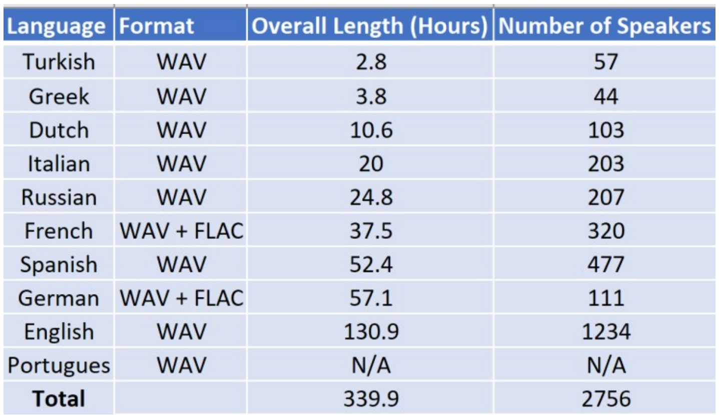
Thus, certain parameters are needed to control the variability of the dataset, to ensure unbiased results. These parameters are also necessary for data selection and refinement. See Table 1.

[44] http://www.voxforge.org/

Table 1: *Data Selection Criteria and Remarks.*

Parameter	Requirement	Method Deployed	Remarks
Magnitude of Dataset (hours)	At least 2.5 hours of speech for each language. Less data would prove difficult to train up statistical models.	The metrics provided on Voxforge was used as a guide, to discard languages with small datasets	Persian (Farsi), Croatian, Hebrew, Albanian, Catalan, Ukrainian, Bulgarian, were dropped from the final dataset.
Number of Speakers	At least 40 speakers. To minimise the probability of the system being speaker-dependent, it is necessary to expose it with as many speakers as possible.	The metrics provided on Voxforge was used as a guide, to discard languages with a small number of speakers	Persian (Farsi), Croatian, Hebrew, Albanian, Catalan, Ukrainian, Bulgarian, were dropped from the final dataset.
Quality of Audio	Discernible speech, with minimal background noise	Random sampling technique – sampled 10% of the audio files of each language. Files with noisy data, singing, intangible utterances were discarded.	The total number of audio files added up to more than 339.9 hours of audio. Thus, it was more time-efficient to use random sampling to check through the datasets.
Sample Rate (kHz)	16 kHz was used, in line with industry standard.	Downloaded 16kHz .wav files directly from Voxforge, and/or passed through ffmpeg for conversion.	See Section 3.3.2

Table 2: *Summary of Dataset*



Language	Format	Overall Length (Hours)	Number of Speakers
Turkish	WAV	2.8	57
Greek	WAV	3.8	44
Dutch	WAV	10.6	103
Italian	WAV	20	203
Russian	WAV	24.8	207
French	WAV + FLAC	37.5	320
Spanish	WAV	52.4	477
German	WAV + FLAC	57.1	111
English	WAV	130.9	1234
Portugues	WAV	N/A	N/A
Total		339.9	2756

'N/A' labels were used to signify missing metric information from Voxforge. However, after downloading and unpacking the datasets, that did not have comprehensive metadata, such as Portuguese, it was found that, with manual inspection, these datasets were sufficient and fitted the selection criteria. Therefore, using the 'total' tally as a minimum count, the dataset included more than 339.9 hours of speech, provided by more than 2756 speakers. This gives rise to a total of 207,463 .wav files.

3.2 Data Assignment

Building such a LID system requires two stages: training and testing. The training stage involves exposing the 'blank' system to a set of data, so the audio files can be analysed, and statistical models are created. The testing stage, then, involves exposing to the system a set of unseen data, which offers a sort of confirmation to verify whether outputs from the models are the expected results. Thus, the data were split accordingly: 80.7% (167,496 .wav files) were assigned as training data; 19.3% (39,976 .wav files) were assigned as unseen, test data. In total, the dataset involves 207, 436 .wav files.

3.3 Data Preparation

3.3.1 Task Grammar

The goal of the system is to identify the language being spoken in an audio file. The system must be able to handle ten languages, without being confined by the content and vocabulary of the utterance. Therefore, a 'blueprint' is needed to inform the system as to what are the possible units that it should recognise and the possible context in which it can appear. It should be noted that by 'units', it means 'languages' in this instance, as the system is trying to represent each language. This 'blueprint' is known as the task grammar.

The Hidden-Markov Tool Kit (HTK) 'provides a grammar definition language for specifying simple task grammars' (Young et al., 2002, p.25). Importantly, the task grammar is directly tied to the acoustic models, thus, the languages encompassed by the WORD variable must be the languages that are used to train the acoustic model.

3.3.2 16kHz Sample Rate and Format Conversion

All data files must comply with HTK's configuration. This involves the conversion of all audio files to .wav format. Moreover, for consistency, the dataset must have a sample rate of 16kHz. 16kHz is an industry -as it is the threshold for which higher levels have no substantial effect on the system performance (Ashihara, 2007).

It is worth noting that some of the data downloaded from Voxforge are corrupted. Some are also saved as older formats, such as FLAC and AIFF. Some of these files encountered problems during the conversion process and thus, they were discarded.

4 System Description

The heart of all speech recognition systems is composed of a set of statistical models, that aims to represent the acoustic and probabilistic information of the dataset. Unlike the static classification methods mentioned above, GMM-HMMs can capture a more comprehensive representation of the audio, as it takes into consideration the entirety of the audio, as opposed to certain preselected components (Muthusamy, Barnard, & Cole, 1994). The following subsections aim to expand on this process in more detail and offer deeper insights into the previously mentioned concepts that are lacking explanation. Importantly, the architecture of this system is an actualisation of the methodologies described by Ambikairajah et al. (2011) and this paper was used as a fundamental guide in the system's blueprint.

4.1 System Design: An Overview

The basic rubrics of a LID system's front-end is illustrated in Figure 1.

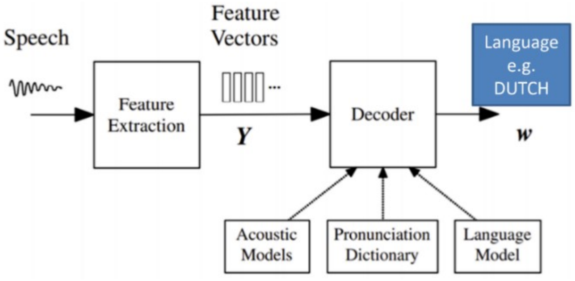


Figure 1: *System Design*

Adapted from (Gales & Young, 2008, p.201) The approach mainly consists of two components: feature extraction and decoding. Fundamentally, these two stages can be conceptualised as 'signal modelling' and 'pattern matching', respectively (Kesarkar & Rao, 2003, p.1).

Broadly speaking, the feature extraction stage involves converting the speech signal to a sequence of acoustic vectors, thus, parameterises the speech waveform. Parameterisation is required for extraction of the most relevant information from speech waveform and disregarding unnecessary noise. In practice, this achieved by converting the speech waveform, frame-by-frame, to a single N-dimensional vector, where *N=26* for this study. Therefore, each utterance is converted into a sequence of vectors, $X = [x_1, x_2, \dots x_{26}]$, where x_{26} is a 26-dimensional vector. The decoder then finds the sequences of words matching to the acoustic vectors by consulting the acoustic models, lexicon, and language model. The latter of which are stored as $\{\lambda_l \mid l = 1,2, \ldots \text{L}\}$, where L is the total number of languages, or in this case, 10. The set of features vectors, *X*, is then used to carry out model training, where a separate model, λ_l is generated for each possible language, *l*.

The feature extraction stage is then repeated for a set of unseen data, in the identification stage. This newly extract feature set is then compared to the model set, $\{\lambda_l \mid l = 1,2, \ldots \text{L}\}$, to identify which *l* has the highest probability of producing the feature vector *X*. Mathematically, this involves pinpointing the language model λ_l that 'maximises a posteriori probability across the set of language models' (Ambikairajah et al., 2011, p. 84). Thus, giving rise to the model below:

(1) $$\hat{l} = arg \max_{1 \leq l \leq L} P(\lambda_l | \boldsymbol{X})$$

Applying Bayes' Rule to (1):

(2) $$\hat{l} = arg \max_{1 \leq l \leq L} \frac{P(\boldsymbol{X}|\lambda_l)P(\lambda_l)}{P(\boldsymbol{X})}$$

Notably, the key assumption made by Ambikairajah et al. (2011) is that each language model has the same likelihood and that, irrespective of the language model, P(X) is the same. The task at hand can thus be thought as identifying the language model that corresponds to the highest probability of *X* occurring (Ambikairajah et al., 2011, p. 84):

(3) $$\hat{l} = arg \max_{1 \leq l \leq L} P(\boldsymbol{X}|\lambda_l)$$

The following subsections will look at these components and stages in greater detail.

4.2 Feature Extraction

The feature extraction technique used by HTK is Mel-Frequency Cepstral Coefficients (MFCCs). In total, there four stages to feature extraction and they are: (1) pre-emphasis, framing and windowing, (2) Fast Fourier Transform, (3) Mel Filter Bank, and (4) Log() Compression and Discrete Cosine Transform.

4.2.1 Pre-emphasis, Framing and Windowing

The first step in MFCC extraction involves boosting the speech signal by passing it through a filter that emphasises higher frequencies. The goal of this step is to 'compensate the high frequencies that are suppressed during humans' sound production', and in turn, amplifying 'the importance of high-frequency formants' (Jang, 2011). As one would expect, there tends to be less energy at the higher frequencies, thus, this boosting is needed to overcome spectral tilt. In this specific scenario, the PREEMCOEF is set to 0.97, to achieve the necessary boost.

Windowing refers to segmenting the speech signal into small blocks of duration. These segments, or windows, are then used to determine each parameter vector (Singh & Rani, 2014). In this case, the window size is 25ms[45]. Importantly, to ensure continuity of the signal, each window overlaps with each other by 10ms. MFCCs are consequently extracted for each window.

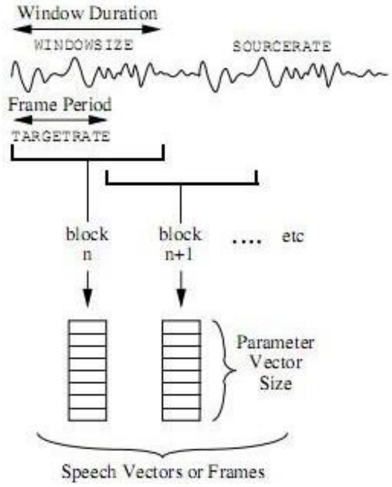


Figure 2: *Visualisation of Windowing* (Young et al., 2002, p.92)

4.2.2 Fast Fourier Transform (FFT)

Fast Fourier Transform is then applied to each frame. The purpose of FFT is for the conversion of the time domain into a frequency domain. Thus, the magnitude frequency response of each frame can be obtained (Singh & Rani, 2014).

4.2.3 Mel-Spaced Filter Bank

The human ear is, arguably, the best speech recognition system currently available. This is accredited to its ability to disentangle arbitrary frequencies across the audio spectrum (Shrawankar & Thakare, 2013). Thus, drawing inspiration from the human ear, the design of MFCC extraction aims to 'operate in a similar non-linear manner, to improve recognition performance' (Young et al., 2002, p.94).

Using HTK's filter bank, that is based on Fourier transform, to produce an equal resolution on a Mel-scale, triangular filters are applied, and they are positioned equally throughout the Melscale (Young et al., 2002, p.95). In this case, a Mel filter bank of 26 channels is applied to the signal, spaced equally throughout. The Mel-scale is used to imitate the non-linear human ear perception of sound, so, it is more discriminative at lower frequencies than higher frequencies (Fayek, 2016). This means that a Mel-spaced filter bank tends to identify fewer extraction points at the top of the spectrum. Taking the windows mentioned previously, each window is transformed using a Fourier transform, and the magnitude values are obtained. Conceptually, this can be thought of as an attempt to capture the energy values of each window, to capture characteristic information of the sound. These energies are known as filter bank energies (FBEs).

[45] HTK specifies all durations in 100 nano second (ns) standard units.

4.2.4 Log() Compression and Discrete Cosine Transform (DCT)

Finally, a Discrete Cosine Transform (DCT) compresses and ranks the values representing the 'power spectrum' to those which are most useful for speech recognition. The log-compressed FBEs' correlations are reduced during this process. Moreover, redundant noise from the speech signal is also refined. For this system, 12 cepstra are returned.

4.3 Decoder

4.3.1 Acoustic Models

In a LID system, the acoustic model represents the acoustic features of a language. Taking the MFCC vectors of the training dataset, acoustic models are constructed for each language, using a 3-state-HMM, with emission distributions in the form of Gaussian Mixture Models (GMMs).

The Hidden Markov Model is used to model sequential data. As implied by the name, much of the underlying data is hidden or remains unknown. This ties in with the main assumption of the approach – that, only the probabilities of the current state are needed to predict the next one. The Hidden-Markov chain used in this project consists of three states and transitions, as well as the probabilities associated with them. More specifically, the HMM is defined by the following set of parameters:

> **Initial state probability $\boldsymbol{\pi}$:** a probability distribution that represents the probability of the HMM starting in each state.
>
> **Transition probability $\boldsymbol{A}$:** a matrix that indicates the likelihood of transition to other states in the HMM, given the current state.
>
> **Emission probability $\boldsymbol{B}$:** a group of multivariate probability distributions (one for each state), that indicates the probability of producing the current state occurring, given the observed sequence of acoustic vectors. Noticeably, the need of a multivariate emission distribution is dependent on the observations' multivariate nature.

It should be noted that two additional states are added, namely, an initial state and a final state are added. This is because they are the entry and exit states, which act as dummy-like, non-emitting states. When one of these states is encountered, it will directly skip to the next one. Fundamentally, 'these non-emitting states serve as the connection terminal for two HMMs, enabling continuous speech recognition' (Kasuriya et al., 2003).

In this system, the emission probability distribution is 'based on a weighted sum of multivariate Gaussian distributions' (Ambikairajah et al., 2011, p.90). To obtain a more characteristic representation of the language, a single Gaussian HMMs is converted to multiple mixture component HMMs. This process of increasing the number of components in a mixture is called 'mixture splitting' (Young et al., 2002, p.200). Conceptually, this can be thought of as using two Gaussians to model the same information as a single one did before, and this act of 'splitting' and break-down into smaller parts, means that there are more components to depict a better picture of the data.

Fundamentally, ‘a Gaussian Mixture is a function that is comprised of several Gaussians, each identified by *k*∈*[1,…, K]*, where *K* is the number of clusters in the dataset’ (Carrasco, 2019).
Applying it to the LID task at hand, a Gaussian Mixture is composed of sets of normal distributions, or Gaussians, denoted by *k*. Each Gaussian *k* is defined the following parameters:

Mean (*μ*) defines its centre or the peak.

Variance (σ^2) defines the spread or width of the distribution curve.

A Gaussian bell curve, therefore, is a probabilistic distribution that expresses the probability of a value occurring.

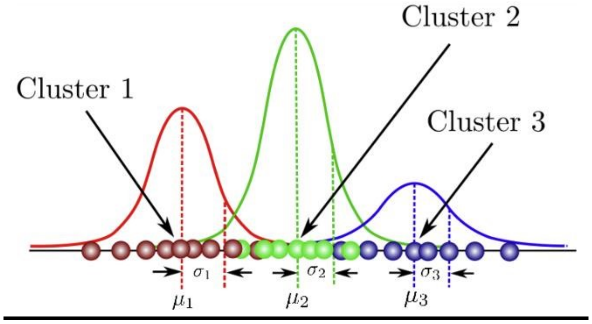


Figure 3: *Gaussians (Carrasco, 2019).*

For the purposes of this system, a more suitable approach is representing the emission distributions as a mixture of multiple multivariate Gaussian densities.

(4) $$g(x) = (\sum_{k=1}^{N} \lambda_k) N(x; \mu_k, \Sigma_k)$$

Or a more robust representation:

(5) $$b_m\left(x^{(t)}\right) = (\sum_{k=1}^{N} \lambda_k^{(m)})\, N\left(x^{(t)}; \mu_k^{(m)}, \Sigma_k^{(m)}\right)$$

where $x^{(t)}$ is an observable vector at time *t*, whereas $\mu_k{}^{(m)}$ is the mean vector and $\Sigma_k{}^{(m)}$ is the covariance matrix of the k^{th} mixture component of the m^{th} state, and $\lambda^{(}{}_{k}{}^{m)}$ is the mixture weight, constrained in such a way that:

(6) $$\sum_{k=1}^{N} \lambda_k^{(m)} = 1$$

4.3.2 Textual Representation Output

During recognition, clusters of features, that belongs to each model are then formed. Should the features extracted from the unseen, test samples, fall within the decision boundary of the language, the system would output these vectors as, for instance, ‘German’. During this process, the language model, which quantifies all the possible sequences of acoustic features and it indicates which sequences are likely to occur, and which are less, are also consulted. (See Section 4.1).

Overall, a GMM system is very easy to train, as it does not require phonetic labelling nor orthographic transcriptions of the training speech.

5 Results

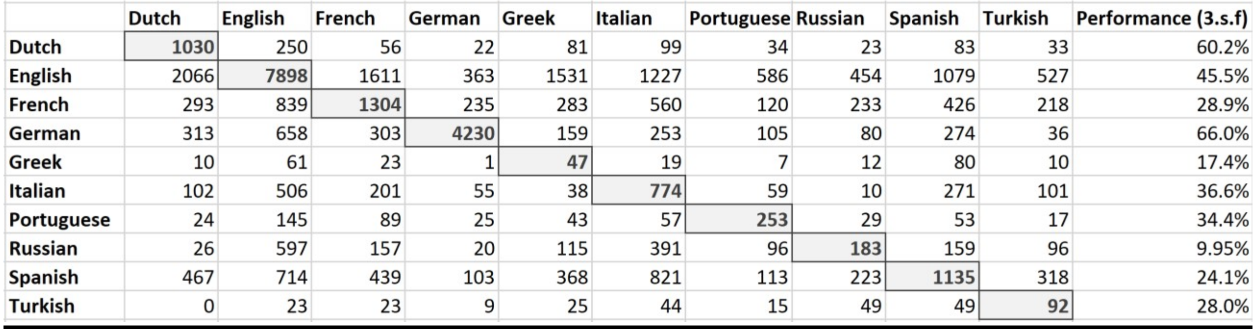
For ease of comparison of the results and further analysis, a way to group the data is needed and this categorisation is significant in decomposing the LID problem into manageable parts. Moreover, many of the recognition patterns also arise from the typological similarities that underpin these languages. Therefore, the 10 languages in the dataset are grouped according to their language family.

Table 3: *Language Family Categorisation.*

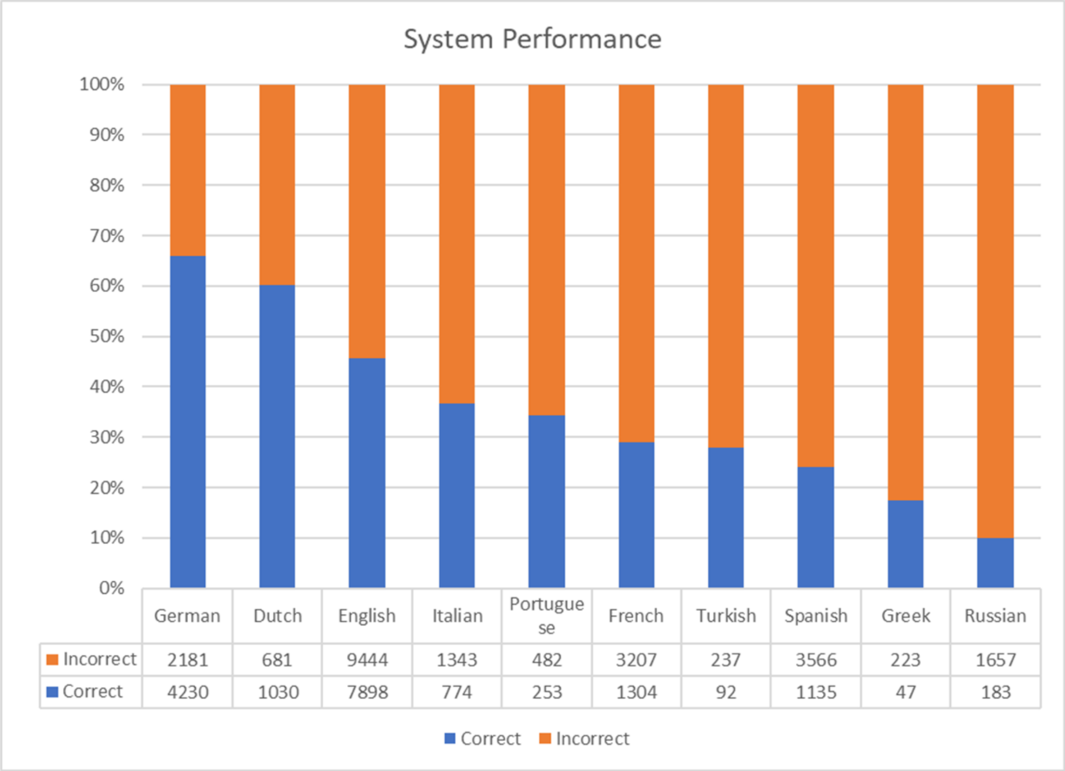
Germanic language(s)	Romance Language(s)	Turkic Language(s)	Slavic Language(s)	Hellenic Language(s)
German	Italian	Turkish	Russian	Greek
Dutch	Portuguese			
English	French			
	Spanish			

The results of the system performance are illustrated below in a confusion matrix. For visualisation purposes, a stacked box plot is also provided.

Table 4: *Confusion Matrix of Recognition Result.*



	Dutch	English	French	German	Greek	Italian	Portuguese	Russian	Spanish	Turkish	Performance (3.s.f)
Dutch	1030	250	56	22	81	99	34	23	83	33	60.2%
English	2066	7898	1611	363	1531	1227	586	454	1079	527	45.5%
French	293	839	1304	235	283	560	120	233	426	218	28.9%
German	313	658	303	4230	159	253	105	80	274	36	66.0%
Greek	10	61	23	1	47	19	7	12	80	10	17.4%
Italian	102	506	201	55	38	774	59	10	271	101	36.6%
Portuguese	24	145	89	25	43	57	253	29	53	17	34.4%
Russian	26	597	157	20	115	391	96	183	159	96	9.95%
Spanish	467	714	439	103	368	821	113	223	1135	318	24.1%
Turkish	0	23	23	9	25	44	15	49	49	92	28.0%



	German	Dutch	English	Italian	Portuguese	French	Turkish	Spanish	Greek	Russian
🟧 Incorrect	2181	681	9444	1343	482	3207	237	3566	223	1657
🟦 Correct	4230	1030	7898	774	253	1304	92	1135	47	183

Figure 5: *Stacked Box Plot of Recognition Results.*

The confusion matrix shows the performance of the system. The overall accuracy of the system is 42.8%. whereas the average performance for a single language is 35.1%.

From the results, it is visible that, the system performed best in identifying German utterances. At a 66.0% correct recognition accuracy, this places German at the top of the ranking for performance, out of the ten languages. Notably, the two languages that follow German in classification performance are Dutch and English. Thus, there seems to be a pattern of the system excelling in the identification of Germanic languages. Dutch, similar to German, also has a high probability of being recognised. 60.2% of the Dutch test files were correctly identified, and 250 (14.6%) recordings were wrongly classified as English. For samples in English, the system correctly predicted 45.5% of the samples to be English. This 14.7% drop in performance, mainly was due to the system misclassifying 2066 (11.9%) files of the English test dataset, like Dutch. On average, the German languages had a classification accuracy of 57.2%.

Romance languages have an overall lower classification accuracy of 31.0%. Out of the four languages, Italian had the lowest error rate, whereas Spanish had the highest. Portuguese and French followed Italian, respectively, but Spanish was preceded by Turkish. Compared to the Germanic languages, the spread of the performance accuracy is smaller, with a standard deviation of 5.63 compared to 10.6.

Interestingly, out of the three languages (Turkish, Greek and Russian), that are 'in isolation', that is, being the only languages representing a language family in the dataset, Turkish had the highest classification rate. With a classification accuracy of 28%, Turkish is ranked 7th preceding Greek (17.5%) and Russian (9.95%). Noticeably, from the Turkish results, equal amounts of data files (2 x 14.9%) were classified as Russian and Spanish, and similarly, for English and French (2 x 6.99%). These equal amounts of misclassification files could indicate the models have picked up similarities across these languages with Turkish, and thus, it was biased towards not classifying these samples as Turkish.

The system performed slightly below chance level for Russian. 'Chance level' is defined as the probability of an event occurring, in the context of an unbiased, random choice (Batanero, Henry, & Parzysz, 2005). Thus, the chance level threshold is 10.0%, as the system has a 1 in 10 chance of achieving correct classification. Visibly, out of the misclassifications, the system misclassified most of the Russian samples (32.4%) as English.

5.1 T-test on Duration

One of the experimental parameters that could impact the system's performance is the duration of the data samples. By taking a random sample of the durations of 40 audio files from each language, Student's t-test was carried out. The suspected correlation is shorter samples are more likely to yield incorrect results, due to the limited acoustic information they carry.

Therefore, a two-sample unequal variance one-tailed test was computed. A result of 0.0858 was obtained and testing at 5% significance level, this suggests that there isn't a significant effect on duration and recognition accuracy. However, this does not eliminate the probability that, there exists no correlation at all between the duration of the data and the accuracy of identification.

6 Discussion

6.1 Phenomenon of Grouping

The approach taken involves scrutinising the patterns and trends visible in different language families. As can be seen from Figure 5, the accuracy of the system forms clusters, centred around languages from the same language family. To explain this phenomenon of 'grouping', i.e., languages in the same family having similar results, the unique acoustic and vocalic features, possessed by these languages are explored. Similarly, there are also features that these languages share, which could be the reason behind misclassification.

6.1.1 West Germanic Languages

The system seems to be able to differentiate German, Dutch and English more accurately than the rest of the dataset. Coincidentally, these three languages are all Germanic languages, and they belong to the same language family. What was particularly interesting to see, apart from the high accuracy rate, was, in fact, the misclassifications the system made. When exposed to German samples, the most common mistake the system made was classifying 10.3% of the data as English. Similarly, when exposed to the English samples, the system classified 11.9% as Dutch. There seems to be a pattern in the system either choosing the correct language or one of its counterparts from the same family. To explain this phenomenon for the Germanic languages in question, it is important to consider the overlap in the phonetic inventory of these languages.

Historically, the West Germanic Languages developed in the regions of Elbe, Rhine-Weser and the North Sea. English is a direct descendent of the North Sea Germanic, with the most striking evolutions are the loss of nasals before voiceless fricatives, palatalisation of /k/ before fronter vowels and /j/, and palatalisation of /g/ before front vowels. Though not as 'deviant' as the changes in English, Dutch too, emerged as a distinct branch of West Germanic language, with its unique development of i-umlaut (Buccini, Moulton & Herzog, 2010). From a phonological perspective, despite these historical changes, the phonemic inventory of these languages is nonetheless similar due to descending from the same ancestor. This similarity could be captured by the MFCCs, during the parameterisation process. Speech waves are longitudinal in nature and differentiation of speech events can be achieved via acoustics according to frequency and amplitude components (Ambikairajah et al., 2011).

The misclassification of 10.3% of the German samples as English can, thus, be offered with an explanation. Some of the phonemes of German, English and Dutch could be potentially identified by the models, and similarity between German and English at an acoustic level is picked up. However, more data and a narrower in-depth study would be required to further prove this.

6.1.2 Romance Languages

Setting aside the phenomenon of overlaps in phonemes across the languages, another potential cause of error in recognition is the potential similarity in the acoustic information that the MFCCs aim to represent. Such information, that could be both advantageous and disadvantageous to task of recognition, manifests as pitch and prosodic patterns.

By comparing the frequency of boundary cues of Catalan, Spanish, South European Portuguese and North European Portuguese, Sónia, d'Imperio, Elordieta, Prieto, and Vigário (2007) found that prosodic breaks are usually marked by a High boundary tone in all five languages. The term 'High boundary tone' (HBT) can be understood to mean the surge in pitch that occurs at the start or end of an intonational phrase (Pierrehumbert, 1980). Despite having this common feature, nuclear pitch accent choice, that is, the accent of the head of the prosodic phrase divides these languages into two sets. More specifically, Portuguese and Italian were found to sustain the pitch after L+H* (low tone + accented high tone) configurations and a continuation rise in pitch is observable for L*+H (accented low tone + high tone) configurations (Sónia et al., 2007). However, the same patterns were not found for Spanish and Catalan.

This could be an implausible explanation as to exactly which features the statistical models have picked out from the speech datasets, but it is an explanation, nonetheless, that sheds light on the common acoustic features present in Italian and Portuguese. It is also plausible in explaining why Spanish had a noticeably lower accuracy rate, with 17.5% of its samples classified as Italian. Again, the idea of the system recognising samples as another 'related' language within the same family, hints at its ability in discerning prosody and intonation.

6.2 Individual Languages

Greek, Turkish and Russian are the three remaining languages in the dataset. They are ‘isolates’, in the sense that, they are the only members in the dataset of a particular language family. Thus, these datasets are more independent, as they are not affected by similar languages.

6.2.1 Turkish

Turkish is the only Turkic language in the dataset, and the system was able to classify 28.0% of the Turkish samples correctly. Phonologically, Turkish has sounds that overlap with Russian and Greek such as the velarized alveolar lateral approximant /ɫ̪/ and voiced palatal plosive /ɟ/ (Yavuz & Balcı, 2011). Yet, the recognition rate of Turkish was still quite high, relative to the other ‘language isolates’. This perhaps is due to prosodic features that mark Turkish apart from the other languages. It would be interesting to see how the classification accuracy would change if another Turkic language were to be introduced. Such a set of results would be able to confirm whether languages from the same family aids identification, or whether it is a hindrance.

According to Torres-Carrasquillo et al. (2002), a GMM is thought to approximate ‘the acoustic phonetic distribution of a language’. Manchala, Prasad, and Janaki (2014) further agrees with this idea, and so, it is believed that ‘each Gaussian density captures some broad phonetic class’ (p. 100).

Strikingly, the system did not identify any of the Turkish samples as Dutch and a small proportion of the samples were identified as German. Thus, from a phonological standpoint, this shows that, the system is perhaps capable of identifying small differences that sets Turkish apart from the Germanic languages. This phonological feature could be final devoicing in Turkish (Hulst & Typology of Languages in Europe (Project), 1999) and the phenomenon of aspirated stops and pre-voiced stops, which are absent from German, Dutch and Russian (Petrova et al., 2006).

6.2.2 Greek

Greek was the smallest dataset out of then 10 languages. Although the size of the dataset does not introduce bias, since the statistical models are trained for each language, having a small dataset means that the models are exposed to limited acoustic characteristics of the language. Thus, the presence of more data would be vital in building a more characteristic model.

6.2.3 Russian

The performance of Russian at chance level could be explained by the methodological practices taken. For this specific dataset, the parameter for mixture splitting was defined as ‘2’. From the results of the preliminary study, it is apparent that this parameter directly affects the classification result for some languages (Mi, 2021). From the initial study, German also had a chance level performance rate, at 3.2%, however, this was changed to 66.0% once the mixture splitting parameter was applied. Theoretically, this means that splitting the mixture into two was beneficial in capture the likely acoustic shape of the signal, for some languages, but not others. Therefore, to obtain the optimal system performance, a set of suitable mixture splitting parameters, such as [16, 32, 64, 128], should be tested to find the best parameter.

6.3 Data Imbalance

An important factor to mention is the existing data imbalance in the study. This is a prominent factor in this system as each individual language is not represented by equal amounts of data. Thus, the performance of one language, such as the likes of the West Germanic languages might seem impressive, but this is also because they have been trained on relatively large amounts of data, whereas the languages with lower recognition accuracy have not. Thus, perhaps direct comparison of the system's accuracy should be completed under more context and scope.

Moreover, the 'metadata' of the dataset remains unclear; many factors, such as age and gender of speakers who contributed the data. This means that the possibility of the system being gender/age- biased cannot be completely ruled out. To eliminate this uncertainty, again, emphasises the importance of a balanced dataset and its necessity for such projects.

6.4 Means of Improvement

This section explores ways in which the overall classification error rate of 57.2% can be reduced. In practice, there are various ways to achieve this, the points explored below are by no means exhaustive.

6.4.1 Variability of Dataset

Speech data is very intricate as it is not consistent like other, more tangible, datasets. This means that speech data samples are varied in nature and this variability directly affects the systems' performance and training result.

The core values of the Voxforge dataset relies on it serving as an open-source speech repository. Practically, users are encouraged to contribute to the project, via submission of their speech. This kind of 'data collection' method is extremely inconsistent, as there is no feasible way in assuring that, all the data is recorded under the same conditions, using the same technical setup. Therefore, the dataset is incredibly varied in quality, both in terms of acoustics and ambient background noise. For some languages, some users have contributed a lot of data to the repository. If the conditions in which they completed the recording is noisy, then this would directly impact the trained acoustic model, and introduce an element of bias.

Furthermore, the absence of users' metadata regarding age and gender means that meaningful comparisons cannot be made. For example, it is not possible to eliminate the probability that, the misclassification of the data could be due to the models being more in conformity with younger, female speakers, who prototypically have higher formant values.
Therefore, to eliminate the possibility of 'gender-dependence', metadata is needed for this dataset.

6.4.2 Use of MFCCs

Phonetically, speech consists of sequences of sound units (phonemes) produced by the excitations of the vocal tract. Each of these phonemes is characterised by sets of formant frequencies, which corresponds to the resonances of the vocal tract (Ogden, 2017). Therefore, formants are one of the most prominent acoustic cues for identification purpose. However, despite this revelation, formant frequencies have seldomly been used for LID.

Consequently, Manchala et al. (2014) proposed a new feature vector, that extracts both the MFCCs of the speech signal and also formant frequencies, through linear prediction analysis. They then proceeded to use this approach in a GMM-based system, and it was found that MFCC features vectors coupled with formant values gave a more comprehensive representation of the acoustic features of the speech signal, which in turn, improved the LID performance (p.104). Therefore, drawing inspiration from Manchala et al. (2014), a similar approach could be taken to improve the performance of this system.

6.4.3 Noise in Dataset

With the goal of increasing the robustness of LID systems in mind, Deshwal, Sangwan and Kumar (2020) investigated the performance of a GMM-based system under conditions of background noise. It was found that the efficiency of such a LID system was strongly affected by noise, as the accuracy of the system with noisy samples decreased by 40%, compared to that of clean data. Furthermore, spectral processing techniques such as Spectral Subtraction (SS) and Minimum Mean Square Error (MMSE) were found to be particularly useful, as they improved the system by 17-20%.

The process of SS involves observing the noise spectrum from periods where the signal, or speech, is absent and only noise is present. Once this estimation of the average noise spectrum is obtained, this is then subtracted from the overall signal, compensating the magnitude of the speech signal (Vaseghi, 1996). MMSE method, on the other hand, is a probabilistic estimator with minimum mean squared errors (meaning it is optimal) (Zhou & Chen, 2015, p.2), and it plays a direct role in the calculation of the vectors during the feature extraction stage. By taking these spectral enhancements and noise suppression methods into consideration, the system can be drastically improved.

6.4.4 Other Speech Information for LID

Hockett (1963) made an empirical generalisation that intonation is universal to all languages, regardless of origins. The term ‘intonation’ can often be founded to appear interchangeably with the term ‘prosody’. For clarity, ‘intonation’ is defined in a broad sense, composing of factors such as word-stress, tone, duration, pitch and intensity.

Often the duration characteristics of phonemes shared across different languages are determined by the different phonetic constraints of the language. All languages utilise pitch to convey surprise, irony or to pose questions. For some languages, tonal pitch variations are crucial in the identification of the language, such is the case for Mandarin Chinese, Thai and Vietnamese. For others, patterns of stress can be used to identify languages that has a word-final stress, such as French, and languages with word-initial pattern, such as Hungarian (Schultz & Kirchhoff, 2006).

From previous research completed on pitch movements in Dutch (Hart & Collier, 1975) and in English (Willems, 1983), it was found that rising and falling pitch movements, i.e., frequency values, are crucial in modelling intonations. This generalisation is also believed to be true for German (Adriaens, 1984). Adriaens (1984) stated that ‘the mean value of the fundamental frequency slowly decreases at a rate which depends on the utterance length’ (p. 37). Adriaens then proceeds to map this as ‘declination lines’. He found that German, like British English, distinguishes between three levels of pitch, whereas Dutch only possesses two. Moreover, it was discovered that German has a significantly steeper declination of pitch than Dutch (Adriaens, 1984).

Therefore, prosody parameters can be taken into consideration for improvements of a LID system. Although the system captures limited prosodic patterns by means of the MFCCs, it does not explicitly try to capture prosodic information, nor does it make the best use of them. To utilise prosodic information, tone can be translated to acoustic parameters, such as pitch or F0, rhyme can be translated into duration sequence and intensity can be used to parameterise stress (Ambikairajah et al., 2011).

7 Conclusion

To conclude, this Independent Study project explored automatic language identification systems. A 10-languages GMM-HMM system was built using HTK. The overall system accuracy was found to be 42.8%. This was attributed to the system misclassifying samples from the same typological language family and experimental shortcomings in the methodology. Ways of improvement have been provided as well as enhancements strategies in dealing with the varied nature of speech data. As a further study, different numbers of mixtures should be tested and exploring the performance of such a system on varying parameters such as distance from microphone, and application of other feature extraction techniques (Mi, 2021).

8 References

Ambikairajah, E., Li, H., Wang, L., Yin, B., & Sethu, V. (2011). Language identification: A tutorial. *IEEE Circuits and Systems Magazine, 11*(2), 82-108.

Ashihara, K. (2007). Hearing thresholds for pure tones above 16 kHz. *The Journal of the Acoustical Society of America, 122*(3), EL52-EL57.

Association, I. P., & Staff, I. P. A. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*: Cambridge University Press.

Baldwin, T., & Lui, M. (2010). *Language identification: The long and the short of the matter.* Paper presented at the Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics.

Batanero, C., Henry, M., & Parzysz, B. (2005). The nature of chance and probability. In *Exploring probability in school* (pp. 15-37): Springer.

Buccini, A. F., Moulton, W. G., & Herzog, M. I. (2010). West Germanic Languages. In *Britannica*.

Carrasco, C. O. (2019). Gaussian Mixture Models Explained. Retrieved from <https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95>

Cimarusti, D., & Ives, R. (1982). *Development of an automatic identification system of spoken languages: Phase I.* Paper presented at the ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Das, H. S., & Roy, P. (2019). Chapter 5 - A Deep Dive Into Deep Learning Techniques for Solving Spoken Language Identification Problems. In N. Dey (Ed.), *Intelligent Speech Signal Processing* (pp. 81-100): Academic Press.

Deshwal, D., Sangwan, P., & Kumar, D. (2020). Language Identification Performance Evaluation Using Spectral Processing. *Available at SSRN 3734808*.

Fayek, H. (2016). Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between. Retrieved from <https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html>

Foil, J. (1986). *Language identification using noisy speech.* Paper presented at the ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition: Now Publishers Inc.

Hockett, C. F. (1963). The problem of universals in language. *Universals of language, 2*, 1-29.

House, A., Neuberg, E., & Wohlford, R. (1975). Preliminaries to the automatic recognition of speech: language identification. *The Journal of the Acoustical Society of America, 57*(S1), S34-S34.

Hulst, H. V. D., & Typology of Languages in Europe (Project). (1999). *Word prosodic systems in the languages of Europe*. Berlin ; New York: Mouton de Gruyter.

Ives, R. (1986). A minimal rule AI expert system for real-time classification of natural spoken languages. *Proc. of 2 nd Artificial Intelligence Advanced Computer Technology*, 337-340.

Jang, J.-S. R. (2011). Audio signal processing and recognition. *Roger Jang's Homepage*.

Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., & Thatphithakkul, N. (2003). *Thai speech corpus for Thai speech recognition.* Paper presented at the Proceedings of Oriental COCOSDA.

Kesarkar, M. P., & Rao, P. (2003). Feature extraction for speech recognition. *Electronic Systems, EE. Dept., IIT Bombay*.

King, S. (2017, 07/11/2017). Re: Filter bank vs. filter coefficients. Retrieved from <https://speech.zone/forums/topic/filter-bank-vs-filter-coefficients/>

Kumar, P., Jakhanwal, N., Bhowmick, A., & Chandra, M. (2011). *Gender classification using pitch and formants.* Paper presented at the Proceedings of the 2011 International Conference on Communication, Computing & Security.

Leonard, R. G., & Doddington, G. R. (1974). *Automatic Language Identification*. Retrieved from

Leonard, R. G., & Doddington, G. R. (1978). *Automatic Language Discrimination*. Retrieved from

Li, K., & Edwards, T. (1980). *Statistical models for automatic language identification.* Paper presented at the ICASSP'80. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Manchala, S., Prasad, V. K., & Janaki, V. (2014). GMM based language identification system using robust features. *International journal of speech technology, 17*(2), 99-105.

Mi, M. (2021). The Unsolved Problem of Language Identification: A GMM-based Approach. Paper presented at the T.W.I.S.T. Conference 2021, Leiden University. Retrieved from <https://conference.studieverenigingtwist.nl/2021/student-speakers>

Muthusamy, Y. K. (1993). A segmental approach to automatic language identification. Citeseer,

Muthusamy, Y. K., Barnard, E., & Cole, R. A. (1994). Reviewing automatic language identification. *IEEE Signal Processing Magazine, 11*(4), 33-41.

Nakagawa, S., Ueda, Y., & Seino, T. (1992). *Speaker-independent, text-independent language identification by HMM.* Paper presented at the Second International Conference on Spoken Language Processing.

Ogden, R. (2017). *Introduction to English Phonetics*: Edinburgh University Press.

Petrova, O., Plapp, R., Ringen, C., & Szentgyörgyi, S. (2006). Voice and aspiration: Evidence from Russian, Hungarian, German, Swedish, and Turkish.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation.* Massachusetts Institute of Technology,

Radha, V. (2012). Speaker independent isolated speech recognition system for Tamil language using HMM. *Procedia Engineering, 30*, 1097-1102.

Rastatter, M. P., McGuire, R. A., Kalinowski, J., & Stuart, A. (1997). Formant frequency characteristics of elderly speakers in contextual speech. *Folia Phoniatrica et Logopaedica, 49*(1), 1-8.

Schultz, T., & Kirchhoff, K. (2006). *Multilingual speech processing*: Elsevier.

Shrawankar, U., & Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. *arXiv preprint arXiv:1305.1145*.

Singh, P. P., & Rani, P. (2014). An approach to extract feature using mfcc. *IOSR Journal of Engineering, 4*(8), 21-25.

Sónia, F., d'Imperio, M., Elordieta, G., Prieto, P., & Vigário, M. (2007). The phonetics and phonology of intonational phrasing in Romance. *Prosodic and Segmental Issues in (Romance) Phonology. John Benjamins (Current Issues in Linguistic Theory)*, 131-153.

Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., & Deller Jr, J. R. (2002). *Approaches to language identification using Gaussian mixture models and shifted delta cepstral features.* Paper presented at the Seventh international conference on spoken language processing.

Vaseghi, S. V. (1996). Spectral subtraction. In *Advanced Signal Processing and Digital Noise Reduction* (pp. 242-260): Springer.

Xu, H., Ding, S., & Watanabe, S. (2019). *Improving end-to-end speech recognition with pronunciation-assisted sub-word modeling.* Paper presented at the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Yavuz, H., & Balcı, A. (2011). Turkish phonology and morphology. *Turkish Phonology and Morphology. Eskisehir: Anadolu Universitesi*.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., . . . Povey, D. (2002). The HTK book. *Cambridge university engineering department, 3*(175), 12.

Zhou, B., & Chen, Q. (2015). A tutorial on Minimum Mean Square Error Estimation.

Zissman, M. A. (1993). *Automatic language identification using Gaussian mixture and hidden Markov models.* Paper presented at the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing.