AUTOMATIC SPOKEN LANGUAGE RECOGNITION FOR MULTILINGUAL SPEECH RESOURCES

[1]MOHAMMED O. ELFAHAL, [2]MOHAMMED E. MUSTAFA, [3]RASHID A. SAEED

Sudan university of Science and technology, Khartoum, Sudan

E-mail: [1]m.osman.h@gmail.com, [2]hafiz85@hotmail.com, [3] eng_rashid@hotmail.com

ABSTRACT

Automatic spoken language recognition refers to a sequence of processes aim to transfer human perception ability of identifying spoken languages to machine using computer program. In spite of great achievements in the domain, the task is still challenging to be practically efficient and reliable. This paper run throughout decades of research attempts approaching optimal languages identification accuracy comparable to human ability of identifying spoken language. Analysis methodologies of extracting most relevant speech information were reviewed. Achievements of approach based on language dependent linguistics rules and those based only on spectral attributes conveys in speech signal were investigated and compared. Exists of standard multilingual speech corpora offers evaluation and comparison of varies speech analysis methods and classification algorithms in single speech variability effects environment. In spite of great achievements, this demanding multilingual communities' communication solution, still looking flexible model easily accepting new language, shorten recognition time, overcoming difficulties of dialects and accents variations and mixed languages speech recognition.

Keywords: *language identification, features extraction, phonotactics, acoustic, mixed speech.*

1. INTRODUCTION

In the universe with 7,099 living spoken languages [1], some of them are spoken only; having no linguistic rules nor orthographic system [2], multilingual communities, either it is physical or virtual, people speaks different languages, dialects and even worse mixed speech in their daily life communication. In such communication, language identification is essential frontend step before any further speech processing step such as routing phone call to human operator fluent in identified language or to language dependent speech instant translator computer application.

The ultimate goal of the process is to transfer most accurate human ability of identifying languages to the machine [3]. Investigating language properties and speech analysis processes and perception mechanism that human uses to identify language in speech utterance are main challenging tasks. Language discriminating attributes uses to determine its identity varies from low level spectral information that speech signal conveys to high level complex linguistics information [3-5]. Language phonemes inventory generated by human articulation system, even it has some common phonemes among languages, considered language dependent information source. Phonotactics, the rules that govern the co-occurrence of speech segments, phonemes, syllable, word, etc., differ from language to other. World languages could be grouped according to prosodic features such as stress, rhythm, duration and intonation. High level linguistic rules of syllable, word and sentence formation for each language is distinctive for languages group to some extent.

Language variations and dialects [6], mixed languages speech [7], spoken only live languages with no orthographic system and linguistic rules [1], raises more challenging identification tasks.

For decades researchers developed techniques, algorithms and data sets for this tightly environment affected task. This paper reviews, orders and categorizes researchers works that sparse technically, environmentally and Chronologically.

The rest of this paper were organized as follows: Section II investigates how human analyze speech signal and what types of speech attributes used to determine language identity. Section III goes through state-of-the-art methods of speech feature extraction and analysis. In section IV, multilingual speech resources and corpora were highlighted. Different approaches that researchers use to accurately identifying languages were reviewed in section V, then important comparison studies and models reviewed in section VI. Section VII discusses and reviews challenges in mixed language speech. The most recent NIST LRE2017 evaluation plan reviewed in section VIII and this work is concluded in section IX.

2. SPOKEN LANGUAGE RECOGNITION

By any means, any oral communication approach between human considered language, either it is language or dialect with orthographic and linguistic rules or not. Automation of human speech perception ability becomes essential for loose world boarders with too many languages to ease communication between its communities either physical or virtual.

2.1 Human Language Perception

Ability of use different language dependent characteristics individually or parallelly make Human the best language recognizer system. Different experiments with linguistic and acoustic language properties shows human fast and accurate ability of identifying his native languages and other languages with little knowledge, beside his good identity judgment for completely unknown languages. Ability of human infant of identifying languages raises the assumption of acoustic signals properties conveys much language dependent information, since infant knows nothing about language linguistic rules [8].

A human speech perceptual experiments were carried with modified speech signal to emphasis desired information on three scenarios: unmodified 6 seconds full speech, prosodic properties destroyed randomly concatenated manually segmented short syllable-like utterance and flattened removed vocal tract information (F0) utterance. The experiments show that prosodic has minimum significant information for language recognition due speech environment variabilities [9]. A threshold based, combined LVCSR and photo-tactic English language detection model investigated against 10 seconds speech from news channel for five languages. This experimental rejection system achieves error rate of 1.8% [9].

2.2 Language Recognition Relative Information

Transferring human ability of language perception to the machine need investigation of language distinctive properties that human uses. The following four main broad categories were used throughout decades of domain researches:

2.2.1 Phonetic inventory

Phoneme is smallest spoken unit that human articulation system produces. Phonetic inventory differs from language to other in term of size of the set, consonants-vowels count and unique-shared phonemes. These inventory properties conclude that even phonemes are shared among languages each language has a unique set of phonemes.

2.2.2 Phonemes co-occurrence

Each language has set of constraints govern the co-occurrence of phonemes called phonotactics. Even inventory set shared among languages the way phonemes structured and ordered differ. For example, د ز are shared phonemes between Arabic and Persian but constrained by their ordered in Arabic language.

2.2.3 Expressive properties

To express meaning, punctuation and sentence structure, speakers change their articulation system configuration to show what called prosody features of stress, duration, intonation and syllable.

2.2.4 Linguistics properties

Each language has phonological rules that govern word formation and syntactic rules for sentence formation.

2.3 General Form of Automatic Language Recognition

Model creation and model testing are two main stages comprised the process of automatic language recognition. Language dependent model is created in training stage using speech sample for each language in the set as illustrated in Figure 1, where n represents languages number in the set.

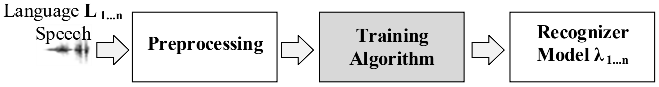

Figure 1: Language recognition training phase block diagram

In the testing stage, the task goes through same process of preprocessing and features extraction along with model created in training stage as illustrated in Figure 2. The selection of language identity decision is taken according to conditional probability as illustrated by Eq. (1).

$$l^{\wedge} = arg_l \max p(l_{1..n} | \boldsymbol{\lambda}_{1..n}) \qquad (1)$$

where: $l^{\wedge}$ target language,
l_n language set, $\boldsymbol{\lambda}_n$ language models

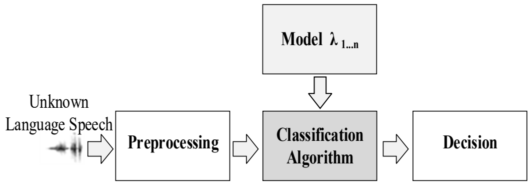

Figure 2: Language recognition testing phase block diagram

3. FEATURES EXTRACTION AND SPEECH PARAMETERIZATION

Essential step in natural speech processing responsible for analyzing speech utterance and extract specific features for further processing. Speech variability due to environment effects, equipment and speakers' status complicates this key process, which it is efficiency impact the overall system performance [10]. Extract task most relevant information and dimensionality reduction are the main goals of speech analysis step.

Preprocessing steps including sampling, signal emphasis to boast desired parts, silence removing which has no related task information, segmenting or framing to suitable stationary part convey enough language information and windowing to smooth frames edge; aims to prepare speech signal to extract desired features that utterance conveys. This process is essential shared step between training and testing stage in languages recognition task. Researchers tries many feature extraction approaches to effectively achieves its goals. The following approaches based on human auditory and perception mechanism were proved to be most effective methodologies.

3.1 Perceptual linear predictive (PLP)

Uses concepts of critical-band spectral resolution, equal-loudness curve and intensity-loudness power law of psychophysics of human hearing process to derive auditory spectrum[11] for further automatic speech processing.

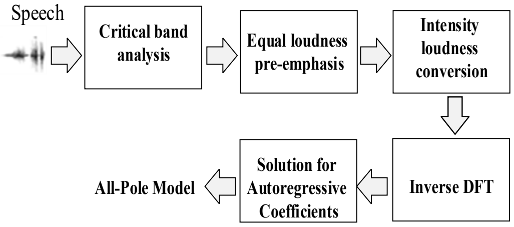
Speech
Critical band analysis
Equal loudness pre-emphasis
Intensity loudness conversion
Inverse DFT
Solution for Autoregressive Coefficients
All-Pole Model

Figure 3: Block diagram of PLP speech analysis

3.2 Mel-Frequency Cepstral Coefficient (MFCC)

is state of the art speech analyzer frontend for varies speech automatic processing backed. MFCC deal with speech utterance logarithmic as human auditory system respond and analyze speech utterance as illustrated by Eq. (2) [12].

$$\mathrm{mel}(f) = 2595 * \log 10 \left(1 + \frac{f}{700}\right) \qquad (2)$$

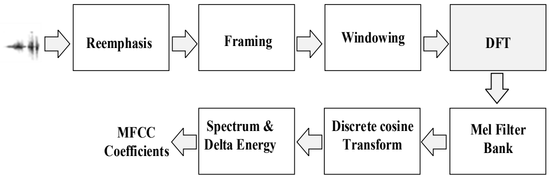

Figure 4: MFCC Block Diagram

3.3 Linear Predictive Coding

The idea is to get correlation coefficients that linearly predicting current sample from previous samples, with error approaching zero. The method encodes spectral envelope (to extract formants) of good quality speech at low bit rate [12].

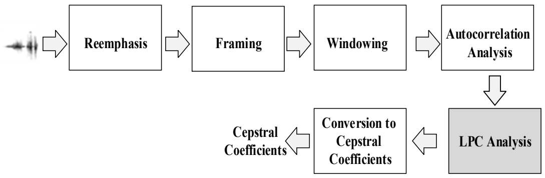

Figure (5): *LPC* Block Diagram

3.4 Joint Factor Analysis (JFA)

To overcome problem of speech sessions variability that relates to speaker condition and capturing environment from recording to other, and approach of decoupling session and language information called Joint factor analysis (JFA) were used, then variability either discarded [13] or modeled beside language useful information. [14]. With universal background model (UBM), is a framework which is a large mixture of Gaussians that covers all speech, and adapted to each language using different techniques such as maximum a posteriori (MAP) algorithms, a JFA adapted language model illustrated by equation (3):

$$m_{data} = M_{ubm} + V_y + D_z + U_{x_{session}} \quad (3)$$

where m are mean supervectors, M is a language independent supervector from UBM, y is the language dependent information assumed to have prior normal distribution, V rectangular matrix of low-rank, D diagonal matrix, z random vector with prior standard distribution, U is an eigenchannel matrix and x is the session variability of the data assumed to be normally distributed. In the training stage, we start training matrix V assuming that D and U are zeros, and then estimate D Given V with assumption U is zero and then find U given estimate of V and D.

3.5 Identity Vector (i-vector)

A data representation (normalization) technique, which is data driven approach that map a supervector (concatenated feature vectors) of sequence of frames for a given utterance (recoding) into informative, fixed length and low-dimensional vector space called total variability space (session and language). This technique reduces mass of data to small manageable and understandable set.

This parametrization strategy motivated by the fact that session/channel factors estimated in JFA also contains speaker and language information, so total variability is modeled by equation (4).

$$M = m + T_w \quad (4)$$

Where m is language and session independent super vector from UBM, T is a rectangular matrix of low rank that defines the total variability space and w is random vector with prior random distribution.

4. MULTILINGUAL SPEECH CORPORA

Human speech environment is highly variable. Speech signal, even for same word in same language, subject to effects of speaker (mode, age, gender and accent), surrounding environment (background noise) and recording equipment configuration and status. This variability makes comparison of models developed in different environment is not applicable and may give misleading results.

Comprehensive development effort of common speech resources were held at Oregon Graduate Institute of Science and Technology in 1993(OGI) [3], two speech corpus where developed, first one contains high quality speech for four languages (American English, Japanese, Mandarin Chinese and Tamil) chosen based on availability of native speakers in United States. The speech automatically segmented using neural network-based segmentation algorithm to vowels fricatives, stops, closures (silence or background noise), pre-vocalic sonorant, inter-vocalic sonorant and post-vocalic sonorant. In the second corpus, more realistic telephone speech collected for ten languages (English, Farsi (Persian), French, German, Korean, Japanese, Mandarin Chinese, Spanish, Tamil and Vietnamese) selected based on linguistics properties and availability of native speakers in United States [15], the corpus then automatically segmented to previously mentioned seven broad phonetic transcription. These two speech corpora were globally available and extensively used for development and comparison and evaluation of models [16, 17].

5. AUTOMATIC LANGUAGE RECOGNITION APPROACHES

Approaches and methods applied to the domain of automatic languages recognition based on extracts and analyses speech attributes that conveys language discriminate information that human use to determine language identify. The task could be grouped into three broad categories as illustrated in Figure 6 were used for the task.

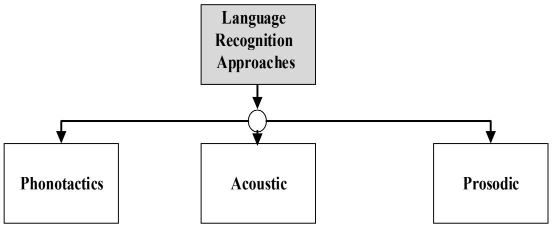

Figure 6: Automatic Language Recognition Approaches

5.1 Phonotactics Approaches

Phonotactics, rules that govern speech formation were used for backend classifier in the automatic language recognition model to specify identity of language. Phonemes and word set, phones co-occurrence, syllable structure and lexical information are examples of such rules. For decades, the approach gives high accurate performance, with shortcomings of needs for large amount of labeled speech data for each language in set, linguistics experts to put rules and relatively long processing time to test rules against incoming speech utterance. The speech tokenizer, a front module to break down speech utterance to smallest units either it is frames, phonemes, syllables or words, is essential frontend part for language recognition Phonotactics approach as illustrated in Figure 7a models creation phase and Figure 7b illustrates model testing phase, and its efficiency affects overall model accuracy [18].

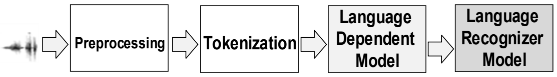

Figure 7a: Phonotactics Approaches Training phase block diagram

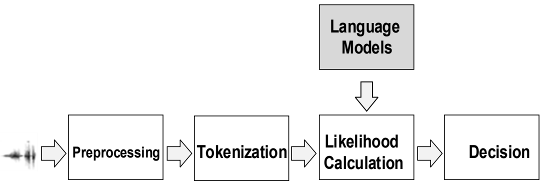

Figure 7b: Phonotactics Approaches Test phase block diagram

In the following section the backend Phonotactics classifier categorized according to level or type of decoder output.

5.1.1 Sequence of Language Key Sounds

Inspired by the fact that each spoken language has a set of distinct sounds (key sounds), early attempt of languages classification for purpose of monitoring communication channels based on language key sounds sequence classification. Both, automatic and manually approach of identifying reference sounds were investigated. Automatic approach gave 64% classification accuracy for seven language, whereas human key sounds preparation approach degrades overall system automation with higher accuracy of 80% for five languages [19, 20]. The success of Large Vocabulary Continuous Speech Recognition (LVCSR) encourage researcher to use more Phonotactics constraint at different tokens level [21-23]. Different algorithms with individual or combined source of information were examined. These studies conclude that using higher linguistics information improve model accuracy and raises the effects of approach drawbacks. Clustering mechanism, based on significant language sounds (key sounds) and sounds co-occurrence used for five Indian languages with VQ to avoid supervised training which is most challenging process in spoken languages recognition. The method achieves promising result on utterance length between 100 and 150ms [24].

5.1.2 Phone based Phonotactics

Phone tokenizer followed by n-gram language model (model that statistically computes co-occurrence probability of tokenizer output sequence) were compared in configuration of Gaussian mixture model (GMM) acoustic based classifier with no labeled data, single-language phone tokenizer followed by n-gram language dependent model (PRLM), parallel PRLM; which uses multiple single-language phone recognizers, each trained in a different language in the set; and language-dependent parallel phone tokenizer along with its n-gram model(PPR) [25]. Different experiments, when applicable, were held for both 10 and 45 seconds utterance length. The comparison concludes parallel PRLM obtain high performance with drawbacks of slow processing and needs for labeled data for each language in the set.

A Hybrid Neural networks and Viterbi algorithm phonemes tokenizer employing temporal pattern is used. The study emphasis dependency between ERR of tokenizer and final output, its concludes that less well-trained tokenizer is better than more with poor training [18]. 5-gram language model following broad phonemes tokenizer achieves performance of 93.7% for phone set of 80 member for 6 seconds utterance length [26]. Comparison of Human perception and machine identification is conducted in the same environment shows that for the short utterance, 1.5 – 2 seconds length, the performance of Human and machine both were below theoretical assumption.

Phone Selection by Elimination (PSE), where mutual information used to select best phones set from set of languages and those phones not selected either removed or substituted followed by language model gives 7.58% EER [27] while target-oriented phone tokenizer (TOPT), where a phones' subset that best discriminates between target languages selected from whole recognizer's inventory, gives 9.26% for 30 seconds length [28].

Phone co-occurrence at the frame level using cross-decoder that considered time aligned information along with frequency of occurrence model slightly improve performance of language identification of the Phonotactics approach [29]. Motivated by this result, with assumption of co-occurrence is language specific, approaches of phone n-gram co-occurrences and co-occurrences of phone n-gram improve baseline Phonotactics approach by 16% [30].

Phone recognition following by language model [25] revisited using phone lattices instead of phones sequence. This approach with neural network classifier outperform traditional one-best phone sequence, which produces 2.7% ERR for 30 seconds utterance length for Arabic, English and Spanish languages [31].

Benefits from vector geometrically that measures similarity as a distance between two vectors, unified phone tokenizer output fed to language n-gram model. The language dependent n-gram model victoried token sequence based on the bag_of_sounds concept. This approach is evaluated with National Institute for Standard and Technology Language Recognition Evaluation (NIST LRE) 1996 and proven successful classification with EER of 14.9% [32]. To eliminate need for large amount of labeled data and linguistic experts for phonotactics approach, a general computationally efficient GMM tokenizer based on acoustic characteristics of speech signal followed by language model have been created. The computationally efficient tokenization step is easily expanded to new languages. In a subset of 12 languages from CALLFRIEND corpus [33], this model produces error rate of 17% [34]. Significant improvement achieved of this low-cost approach by incorporating speech signal temporal information (shifted-delta-cepstral SDC) [6]. This language identification technique applied to dialect identification for dialects in Call Friend and Miami corpus. accuracy of 13% and 30% ERR achieved of dialects in Call Friend and Miami corpus respectively [35].

A JFA a front-end to i-vector for 3-gram counts language model with SVM backend shows slight improvement over baseline Phonotactics model which indicates higher order of n-gram models most probably gives further improvement with less computation cost [36].

5.1.3 Syllable based Phonotactics

Inspiring by the motivated result of preliminary experiment for eight languages, manually broad transcription (stop, fricative, vowel, silence) fed to Hidden Markov Model HMM to model sequential and co-occurrence properties of speech patterns [37], syllables segments for five languages representing two languages families achieved 80% classification accuracy [38], with real male read speech. The study shows syllable perfectly differentiate between two languages family. Automatic segmentation of speech signal based on fundamental frequency (F0) and temporal trajectory of short-term-energy output broadly categorized to Vowels, diphthongs, glides, schwa, stops, nasal, fricatives, and flaps. This frame by frame segment fed to language dependent trigram model of 12-CallFriend languages corpus. The trigram model had 24% ERR for 30 seconds utterance length. the Study concludes that prosodic information is significant in classifying some languages such as mandarin Chinese [39]. With assumption that even shared phones and co-occurrence spread over languages, sound duration is different based on language, context and speaker. Automatic normalized duration vector of UV (Unvoiced, Voiced) segments front-end for n-gram language model achieves 19.7% ERR on NIST LRE 2005 [40]. With Prosodic Attribute Model (PAM), attempt is held to model language-specific co-occurrence of compact prosodic attributes.

Since single language dialects most probably share phonetic inventory and syllable structure and with the same written script, syllable tokens fed to n-gram model along with Latent Semantic Analysis (LSA) to capture more phototactic constraints. For three Chinese dialects (Mandarin, Cantonese, Shanghai) 99.23% classification accuracy where achieved [41].

5.2 Acoustic Approaches

In spite of employing higher linguistics information for automatic languages recognition achieved most identification accuracy; computation complexity and linguistics experts' dependency force researchers looking language dependent information conveys into speech waveform that human with linguistics knowledge or not uses to identify utterance language [42, 43].

Based on infant ability to discriminates between language with no previous linguistics knowledge; French native speakers discriminates well between two different unknown languages having different rhythms using rhythm prosodic property [8].

Acoustic approach looking solution of model expansion limitation on linguistics base approach (easy adding new languages to the system without need of linguistic experts and more training data). Most Phonotactics approach modeled at most 20 and less languages, which represents very small set of common live languages. Rare and detectable languages features are significant regional discriminant languages properties. Features of occurrence of nasalized vowels, labial-velar stops and of retroflex consonants were examined against The University of California, Los Angeles (UCLA) Phonological Segment Inventory Database (UPSID) of 451 languages representing all languages families with at least one language for each family[44]. These three features together eliminates set members to only three languages which represents 0.7% of languages from the corpus [45].

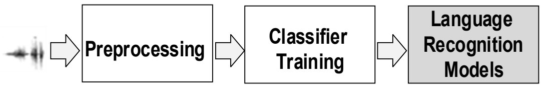

Figure 8a: Acoustic Model Training Phase Block Diagram

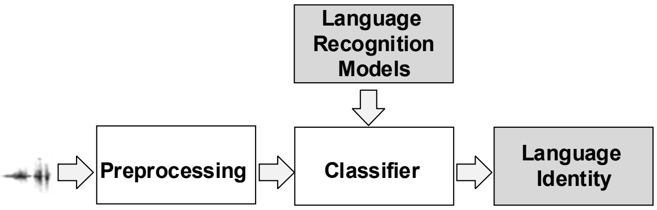

Figure 8b: Acoustic Model Testing Phase Block Diagram

5.2.1 Spectral information based

This approach based only on spectral attributes contained in speech signal.

100 acoustic features, including Autocorrelation, cepstral, filter and formant frequencies, for 30ms frame length speech extracted using Linear Predictive Coding (LPC). For eight languages, three minutes of read speech from adult male randomly divided to test and training part for polynomial classifier achieved 84% overall classification accuracy [43]. The corpus were enlarge to 50 hours of read speech from 122 male for expert system achieved over all accuracy of 92% using 50 distinct language features, extracted by domain experts [46].

A comparison study with configuration for PLP acoustic features from broad and fine auto segmentation of speech utterance followed by feedforward neural network classifier, shows broad segmentation gives accuracy of 70% which is less by 13% from accuracy of fine segmentation with less computation cost for 13.4 seconds average speech length [47]. Employing geometric and algebraic features of vectors; based on similarity measurement between each vectors in destinations vectors set and incoming phone call vector, a domain-independent call routing model routes a caller call to appropriate destination that its vector close enough to incoming phone call vector [48].

Promising SDC technique used in [6] with polynomial expansion tokenizer adopted for specially designed "one vs. all" SVM with sequence kernel (Generalized linear discriminant sequence (GLDS)) that map features of utterances to higher dimension to ease linear separation. Only acoustic features utilized for dataset of 12 languages in NIST LRE2003 evaluation data from CallFreind Corpus. For classifying unseen utterance backend postprocessor pick a language with high score from parallel language dependent SVM model. The result is comparable to GMM pure acoustic models [49] [50] [51]. Based on the fact that human can makes reasonable judgement for unknow language using only acoustic features; a novel approach seek and determine a language specific information significant for language recognition (Perceptually Significant Regions (PSR) in speech utterance) were introduced [52]. Recurrent neural network trained with PSR achieves 9% performance enhancement over other training approach.

With SVM classifier, i-vector technique outperform direct JFA state-of-the art model [53],acoustic model tested on NIST LRE 2009 shows promising results [54]. Double reduction of SDC acoustic speech feature with deep neural network bottleneck feature (feature from layer with few hidden nodes) followed by i-vector representation shows significant improvement with low computation cost for LRE2009 tests of 30, 10 and 3 seconds, specifically for short duration, which achieves 9.71% EER [55].. Inspired by the success of artificial neural network in acoustic modeling, it is used in language identification task as front-end, backend or both with or without total variability modeling. feed forward neural network with deep learning (DNN), that fed of PLP frame-based feature, were used for language identification. The performance outperforms state-of-the-art i-vector approach and achieves 70% improvement on 3s utterance length, specifically when large amount of training data is available. (since i-vector saturated DNN stay learning) [56].

Single feature extractor and i-vector with bottleneck deep neural network for speaker and language recognition, shows good performance and robustness encourage to use such single front-end to develop application for both domain [57]. A pre-trained deep neural network for automatic speech recognition task used to extract i-vector representation for building generalized language identification model with attention to within phone transitions. This approach tested for English, mandarin and Arabic dialects shows good performance and generalization capability [58].

5.2.2 Speech Token based Identification

Inspired by different languages have different phone sets, language with maximum likelihood score were selected as target language from parallel output of HMM language-dependent phone recognizer [59, 60]. The baseline system is for English and French under laboratory conditions, then model extended to use publicly available phone labeled OGI-10 spoken languages telephone speech corpus [15]. The overall classification accuracy for two languages is 82% and 59.7% respectively.

A two stages approach: automatic broad and fine phonetic segmentation followed by classification in second stage [47] for English and Japanese languages were compared. For each stage feedforward neural network trained with backpropagation using PLP features were i. two approaches achieved accuracy of 86.3% and 83.2% respectively for 13.4 seconds average speech length. Inspired by work in [47] a phonemes superset, "instead of separate phonemic front end of each language", of English, Japanese and German languages has been created from two linguistic categories: mono-phoneme subset of little or no overlap between languages and poly-phoneme subset of share phonemes among languages [61]. This study concentrates on mono-phonemes which convey most language discriminant information. For three languages, a neural network classifier for phone-based mono-phonemes superset system achieves over all accuracy of 91% for 45 seconds speech length and 71% for 10 seconds speech length [62]. The study shows that using mon-phoneme superset reduce feature space with insignificant performance loss.

Motivated by difference of number of vowels and its articulation process setup in each language; auto-detection phone recognizer front-end of vowels inventory for five languages (French, Japanese, Korean, Spanish, Vietnamese) from OGI corpus, for likelihood and Euclidean distance classification model, achieves promising performance of 61% for 45 seconds length utterance [63].

With assumption, languages sounds can be collectively clustered by Acoustic Segment Models (ASM) [64] ;where vector of acoustic units along with co-occurrence attributes were formed (high occurrence represents key term), vector space modeling (VSM); the dominant technique in information retrieval (IR) research, were used for language recognition employing this unsupervised approach. VSM discriminately measures the similarity between test (query) vector and target language vector based on distance between them. With SVM classifier the approach achieves EER of 2.75% and 4.02% in 30-s 1996 and 2003 NIST LRE tasks [65]. A Target-Oriented Phone Tokenizer (TOPT), a concept of selecting a subset of phones that best discriminates between languages, were used for VSM backed classifier. Different approaches for driving such set were investigated. The study shows that extracting those significant phones from universal phone recognizer is effective than from language dependent than from parallel phone recognizers. This approach achieves 1.27%, 1.42% and 2.73% EER on the NIST 1996, 2003 and 2007 LRE respectively for 30 seconds length test utterance [28]. As a front-end of VSM classifier, a data-driven technique was used to build universal acoustic tokens based on manner and place of articulation. preliminary result of this ongoing work shows promising improvement of language recognition performance [66].

5.3 Prosodic Information Based

Prosody is a study of tune and rhythm and how they contribute in speech meaning. It is characterized by vocal pitch (fundamental frequency), loudness (acoustic intensity) and rhythm (phoneme and syllable duration).

Human perception study shows that a simple structure prosodic speech features, rhythmic and into national (e.g. fundamental frequency (F0), F0 gradient, intensity and duration) playing significant roles of human process of identifying spoken languages in spite its subject to speech variability such as speakers emotional status [8].

With believe that syllable conveys prosodic features, syllable-like tokenizer was applied. Using mutual information criterion to select and analyze language recognition relative prosody features, claimed to be the best language recognizer model among all prosodic features based models [67]. Formant values and location for 4.5 noisy speech were used to capture sound pattern of three languages from different languages families. Frequency of occurrence of this pattern then clustered using K-means and vector quantization (VQ) algorithm. Variety of rhythm and intonation from language to other also modeled to achieve 39% clustering accuracy [68]. Using LPC to extract formant information from noisy speech greatly enhance the efficiency of this model [69].

Inspired by perceptual and algorithmic experiments shows that prosodic properties (rhythm, stress and intonation) of speech conveys significant language discriminant characteristics, with assumption of phones and their co-occurrence spread over languages, sound duration (rhythm) is different based on language, context and speaker; Automatic normalized duration vector of UV (Unvoiced, Voiced) segments front-end for n-gram language model achieves 19.7% ERR on NIST LRE 2005 [70]. Based on articulation or speech production mechanism, features that relevant to underline language were extracted for each consonant or vowel segment along with syllable supra-segmental and sequence structure properties. Likelihood measure for five Indian languages achieves 65% classification accuracy [71]. The jitter (variability of F0) and shimmer (amplitude of vibration) as new information source were used for SVM classifier with radial basis function kernel. For five Indian languages performance accuracy is 81.4%, 76% and 87.4% for vowel, syllable and word based respectively [72].

Motivated by languages could be grouped into a rhythmic class, five European languages (English, French, Germany, Italian and Spanish) shared two rhythmic families, their pseudo-syllabic extracted automatically using Gaussian Mixture Model (GMM) produce identification performance of 81% for 20 seconds length utterance. Even experiments done on limited dataset, the result shows promising efficiency of unsupervised approach to automatic language recognition [73]. This approach extended to catch up intonation information through fundamental frequency for better separation between languages classes [74].

In spite of difficulties modeling and extracting rhythm, vowel detection algorithm (vowel, non-vowel segment) is used to extract syllable related rhythm (pseudo-syllabic). For seven languages (English, French, German, Italian, Japanese, Mandarin, Spanish) from three rhythmic families. A rhythmic classes clustering achieves 86% - 92% accuracy. For 21 second utterance length for each language, 67% – 75% accuracy were achieved. The later experiment shows that confusion occur for languages from same rhythmic family than other [75]. Inspired by the fact that syllable is more distinctive than phoneme among languages, unsupervised (grouping) syllable tokenizer approach was proposed. The log-likelihood classifier classifies syllable-like tokens with 69.5% 75.9% for 30 and 10 seconds utterance respectively, whereas language distinctive syllables achieves 64.5% and 67.2% receptively for the same test duration from OGI-TS speech corpus [76]. For the prosodic GMM features (rhythm, stress and intonation) is evaluated for i-vector reduced feature space. The result shows fusion i-vector prosodic model with new techniques gives comparable performance to acoustic Phonotactics model [77].

6. EVALUATION AND COMPARISON STUDIES

Comparison and evaluation of automatic languages recognition research output is impractical in different environment due speech variability.

Setting up single acoustic based environment for four languages (English, Japanese, Mandarin Chinese and Indonesian), gives opportunity to compare and evaluate recognition accuracy of vector quantization (VQ), discrete and continuous HMM and Gaussian Mixture Model (GMM) algorithms. Capturing dynamic speech features with LPC and static features with Mel-Cepstrum achieves recognition accuracy of 77.4%, 47.6, 86.3% and 81.1% respectively [78].

A comparison of using acoustic features only and three others setup of Phonotactics approach based on phone recognition followed by n-grams language model concludes that parallel combined process of acoustic and Phonotactics information for languages with enough training data is state-of-the-art language recognition performance at that time [25].

In 1996 NIST begins publish common evaluation environment including speech corpus and test plan [16]. Since then evaluation and competition held every two years, for year 2017 (LRE17) eighth Language recognition evaluation plan is for language detection for 5 languages clusters (Arabic, Chinese, English, Slavic and Iberian) with 14 languages [17]. Motivated by segmentation approach introduced in [37] single independent phone front-end used for multilingual recognition system that uses phonetic, prosodic and Phonotactics information individually and combined together [79]. This study concludes that acoustic based model outperforms language model for short utterance which is contains less linguistic information.

Automatic recognition of Language in speech utterance for languages from same families or that shares many sounds are confusable and add another complexity dimension of the task [80]. For NIST LRE 2009 tasks that includes language identification, target language detection and discriminate between confusable language pairs, MIT Lincoln laboratory submit three systems for 23 languages (Amharic, Bosnian, Cantonese, Creole, Croatian, Dari, English-American, English-Indian, Farsi, French, Georgian, Hausa, Hindi, Korean, Mandarin, Pashto, Portuguese, Russian, Spanish, Turkish, Ukrainian, Urdu, and Vietnamese). A fusion of three recognizers, GMM spectral system (GMM-MMI), SVM GMM super vector spectral system SVM-GSV and SVM language classifier, achieves average 1.64% EER for the identification, detection along with language pair discrimination for 3, 10 and 30 seconds utterance length [81].

The task of Albayzin2012 Language Recognition Evaluation (LRE), effort made by the Spanish/Portuguese community for benchmarking language recognition technology, is to output likelihood scores for the YouTube extracted audio for each target languages (English, Portuguese, Basque, Catalan, Galician and Spanish) along with score for out-of-set languages (French, German, Greek and Italian) that have no training data. State-of-the-art total variability (i-vector) model mostly used for participants submissions [82].

Pear in mind issues of short utterance recognition and linguistics processing content requirements for language identification task, frame by frame identification method were investigated for real time application with deep neural network. For 3 seconds task of NIST LRE 2009 (8 class), for comparison with standard, this method outperform i-vector state-of-the-art by 40% and by 76% using Google 5M LID (34 class) speech corpus for real time testing, because i-vector performance degrades against size of data used to derive it [83]. Convolution deep neural network for 3 seconds utterance length shows comparable performance to i-vector state-of-the-art with reduction of parameters by factor 100 for NIST LRE2009 8 languages [84]. Exploring its ability of store information from previous inputs during long time periods, Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) outperform i-vector and DNNs state-of-the-art approaches of language identification task from short utterance (3 seconds) by 26% performance enhancement for NIST LRE2009 8 languages task, very short utterance till 0.01 second were tested, and 50% accuracy achieve for only 0.5 second utterance length. The system shows robustness and detection of out-of-set and unseen languages [85]. With attention mechanism LSTM RRNs achieves 34.33% performance enhancement over traditional i-vector approach, and 8.2% ERR reduction in compare to frame level LSTM RNNs for 14 languages from NIST LRE2007 evaluation task for short utterance (3s) [86].

Google 5M LID and NIST LRE2009 corpus were used with large languages count beside out-of-set languages. Compact and low-dimensional acoustic feature vectors were generated to model total variability using bottleneck neural network. Fused with Phonotactics the system shows improvement over state-of-the-art system for short duration, testing with NIST LRE2009 achieves EER 1.08%,1.89%, 7.0.1% for 30, 10 and 3 seconds utterance length respectively [87]. Acoustic Deep learning neural network model with bottleneck feature and total variability (i-vector) shows promising language identification result for highly noisy speech [88].

7. NIST LRE2017 Language Recognition Evaluation

Speech environment and speakers' variability makes language recognition techniques comparison not applicable under different testing environments, such as speech capturing equipment, speakers' accent, age and mode and language linguistic formation. To overcome environment variability issue NIST continue offers comparison and evaluation environment to explore new promising techniques, supporting implementation of that techniques and measures its performance in single test environment.

LRE2017 focus on language detection (given a speech utterance and a target language, automatically determine if the target language was spoken in the test utterance) of closely related languages and measures languages similarity confusion, besides effects of amount of training data of system performance, for 3s, 10s and 30s speech utterance length from conversational telephone speech (CTS) and broadcast narrow band speech (BNBS), speech extracted from videos or video speech (VS) speech corpora [17]. The plan targeted 14 languages from 5 languages clusters as listed in Table 1.

Table 1: Target Languages and Clusters

Language Cluster	Target Languages
Arabic	Egyptian Arabic, Iraqi Arabic, Levantine Arabic, Maghrebi Arabi
Chinese	Mandarin, Min Nan
English	British English, General American English
Slavic	Polish, Russian
Iberian	Caribbean Spanish, European Spanish, Latin American Continental Spanish, Brazilian Portuguese

8. AUTOMATIC LANGUAGES RECOGNITION CHALLENGES

In spite of great achievements obtained in this demanding front-end module in multilingual communities' communication; performance is far from human based line system. Apparently, some domain performance challenges arise from speaker and speech environment and others from linguistics structure of languages.

8.1 Identifying Unseen Languages

In spite of good performance achieved of phonotactics based language recognition approach, needs for linguistic experts and big amount of labeled training data slowdown its progress due to model limitation in term of adding new languages to the model and classifying unseen language of world with a lot of language and dialects, some of them are spoken only.

8.2 Language Recognition Time

Fast and accurate automation of language recognition are ultimate goals of the domain, but it is still far away comparable to human performance, the tradeoff between recognition accuracy and recognition time is hot research issue. Some real time speech processing system need recognition time comparable to human performance such as instant translation bearing in mind the task is just pre-process of translation main task, while others systems concentrates on system accuracy such as speech biometrics used for access control.

8.3 Dialects and Accents Variations

Most challenging issue in the domain of speech processing in general and language recognition specifically is language variations (dialects) and speakers' accents difference. For accents variation techniques of total variability that separates language attributes form channel attributes reduce its effects in performance. Unseen language dialects greatly affect the performance of the system. To some extent, this challenge similar to problem of affect adding new languages to the system.

8.4 Mixed Speech Conversation

Bilingual, has great effect in daily life communication, and may represents half population of the world [8]. This add practical complexity dimension of speech processing enabled system. A practical use of spoken language recognition for mixed speech[7], is applied as a front end for multilingual speech recognition system of English and mandarin languages. The study shows that performance of speech recognition system in this environment greatly enhanced by perfect language recognition using less than 3 second utterance length before speech recognition process start [89].

9. CONCLUSION

For decades, researchers investigate varies approaches and techniques heading ultimate goals to fast and accurately identifying language in speech utterance based on human ability of such task.

Variability of speech and effects of its environment complicates the process of extracting most relevant language information. develop efficient and best fit training algorithms with minimum needs for linguistics experts still challenging task. Language dialects, mixed languages speech and languages with no linguistics rules add other dimensions of task complexity.

Exists of standard multilingual corpora, greatly enhanced model performance by offer single evaluation and comparison environment of varies techniques.

REFERENCES

[1] G. Simons and C. D. Fennig. (2017, 22/10/2017). *Ethnologue: Languages of the World, Dallas, Texas: SIL International (20th ed.).* Available: https://www.ethnologue.com/ethnoblog/gary-simons/welcome-20th-edition

[2] G. Simons and C. D. Fennig. (2017, 22/10/2017). Ethnologue: Languages of the World (Aka, the language of Sudan), Dallas, Texas: SIL International (20th ed.). Available: https://www.ethnologue.com/language/soh

[3] Y. K. Muthusamy, "A segmental approach to automatic language identification, PhD Thesis," Oregon Graduate Institue of Science and Technology, October 1993.

[4] K. M. B. Marc A. Zissman, "Automatic Language Identification," *speech comunication,* vol. 35, pp. 115-124, 2001.

[5] B. M. Haizhou Li, Kong Aik Lee, "Spoken Language Recognition: from Fundamentals to Practice," *in proc. IEEE Spoken Language Recognition,* vol. 101, pp. 1136 - 1159, 2013.

[6] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using Gaussian mixture models

and shifted delta cepstral features," in *Interspeech*, 2002.

[7] C.-H. Wu, Y.-H. Chiu, C.-J. Shia, and C.-Y. Lin, "Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs," *IEEE Transactions on audio, speech, and language processing,* vol. 14, pp. 266-276, 2006.

[8] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: A study based on speech resynthesis," *The Journal of the Acoustical Society of America,* vol. 105, pp. 512-521, 1999.

[9] J. Navratil, "Spoken language recognition-a step toward multilinguality in speech processing," *IEEE Transactions on Speech and Audio Processing,* vol. 9, pp. 678-685, 2001.

[10] F. Verdet, "Exploring variabilities through factor analysis in automatic acoustic language recognition," PhD Thesis, Faculty of Science, University of Fribourg, Switzerland), 2011.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America,* vol. 87, pp. 1738-1752, 1990.

[12] E. M. Mohammed, M. S. Sayed, A. M. Moselhy, and A. A. Abdelnaiem, "LPC and MFCC performance evaluation with artificial neural network for spoken language identification."

[13] F. Verdet, D. Matrouf, J.-F. Bonastre, and J. Hennebert, "Factor analysis and svm for language recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[14] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, pp. 788-798, 2011.

[15] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The Ogi multi-language telephone speech corpus," in *ICSLP*, 1992, pp. 895-898.

[16] NIST. (2017, 9/10/2017). *Language Recognition Evaluation Plans*. Available: https://www.nist.gov/itl/iad/mig/language-recognition

[17] NIST. (2017, 9/10/2017). *NIST 2017 Language Recognition Evaluation* Available: https://www.nist.gov/itl/iad/mig/nist-2017-language-recognition-evaluation

[18] P. Matejka, P. Schwarz, J. Cernocký, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Interspeech*, September 2005, pp. 2237-2240.

[19] R. G. Leonard and G. R. Doddington, "Automatic Language Identification," Air Force Rome Air Development Center Technical Report RADC-TR-74-200, August 1974.

[20] R. G. Leonard, "Language Recognition Test and Evaluation," Air Force Rome Air Development Center Technical Report RADCTR-80-83, March 1980.

[21] S. Mendoza, L. Gillick, Y. Ito, S. Lowe, and M. Newman, "Automatic language identification using large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, pp. 785-788.

[22] T. Schultz, I. Rogina, and A. Waibel, "Experiments with LVCSR based language identification," in *proc. ICASSP*, 1995.

[23] T. Schultz, I. Rogina, and A. Waibel, "LVCSR-based language identification," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, pp. 781-784.

[24] J. Balleda, H. A. Murthy, and T. Nagarajan, "Language identification from short segments of speech," in *INTERSPEECH*, 2000, pp. 1033-1036.

[25] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing,* vol. 4, p. 31, 1996.

[26] M. Adda-Decker, F. Antoine, I. Vasilescu, L. Lamel, J. Vaissiere, E. Geoffrois, *et al.*, "Phonetic knowledge, phonotactics and perceptual validation for automatic language identification," in *In ICPhS*, 2003.

[27] C. S. Kumar, H. Li, R. Tong, P. Matějka, L. Burget, and J. Černocký, "Tuning phone decoders for language identification," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 5010-5013.

[28] R. Tong, B. Ma, H. Li, and E. S. Chng, "A target-oriented phonotactic front-end for spoken language recognition," *IEEE transactions on audio, speech, and language processing,* vol. 17, pp. 1335-1347, 2009.

[29] M. Penagarikano, A. Varona, L. J. Rodríguez-Fuentes, and G. Bordel, "Using cross-decoder phone coocurrences in phonotactic language recognition," in *Acoustics Speech and Signal*

Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 5034-5037.

[30] M. Penagarikano, A. Varona, L. J. Rodríguez-Fuentes, and G. Bordel, "Improved modeling of cross-decoder phone co-occurrences in SVM-based phonotactic language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2348-2363, 2011.

[31] J.-L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," in *INTERSPEECH*, 2004.

[32] H. Li and B. Ma, "A phonotactic language model for spoken language identification," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 515-522.

[33] Linguistic Data Consortium. (1996, 22/10/2017). *CALLFRIEND Speech Corpus*. Available: http://www.ldc.upenn.edu/ ldc/ about/callfriend.html

[34] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller, "Language identification using Gaussian mixture model tokenization," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, 2002, pp. I-757-I-760.

[35] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian mixture models," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[36] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen, "iVector approach to phonotactic language recognition," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[37] A. S. House and E. P. Neuburg, "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations," *The Journal of the Acoustical Society of America*, vol. 62, pp. 708-713, 1977.

[38] K. Li and T. Edwards, "Statistical models for automatic language identification," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80.*, 1980, pp. 884-887.

[39] A. G. Adami and H. Hermansky, "Segmentation of speech for speaker and language recognition," in *INTERSPEECH*, 2003.

[40] B. Yin, E. Ambikairajah, and F. Chen, "Voiced/unvoiced pattern-based duration modeling for language identification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4341-4344.

[41] B. P. Lim, H. Li, and B. Ma, "Using local & global phonotactic features in Chinese dialect identification," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, pp. I/577-I/580 Vol. 1.

[42] H. Combrinck and E. Botha, "Automatic Language Identification: Performance vs. Complexity," in *In Proceedings of the Sixth Annual South Africa Workshop on Pattern Recognition*, 1997.

[43] D. Cimarusti and R. Ives, "Development of an automatic identification system of spoken languages: Phase I," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, 1982, pp. 1661-1663.

[44] Ian Maddieson and K. Precoda. (1984, 11/10/2017). *UCLA Phonological Segment Inventory Database (UPSID)*. Available: http://phonetics.linguistics.ucla.edu/sales/software.htm#upsid

[45] J.-M. Hombert and I. Maddieson, "The Use of'Rare'Segments for Language Identification," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[46] R. Ives, "A minimal rule AI expert system for real-time classification of natural spoken languages," in *2nd Annual Artifiial Intelligence and Advanced Computer Technology Confernce*, Long Beach, CA, may 1986.

[47] Y. K. Muthusamy, K. M. Berkling, T. Arai, R. A. Cole, and E. Barnard, "A comparison of approaches to automatic language identification using telephone speech," in *EUROSPEECH*, 1993, pp. 1307-1310.

[48] J. Chu-Carroll and B. Carpenter, "Vector-based natural language call routing," *Computational linguistics,* vol. 25, pp. 361-388, 1999.

[49] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, and D. A. Reynolds, "Language recognition with support vector machines," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[50] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-

Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language,* vol. 20, pp. 210-229, 2006.

[51] K.-A. Lee, C. You, and H. Li, "Spoken language recognition using support vector machines with generative front-end," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4153-4156.

[52] J. Braun and H. Levkowitz, "Automatic language identification with perceptually guided training and recurrent neural networks," in *International Conference of Spoken Language Process*, Sydney, Australia, 1998, pp. 289–292.

[53] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Proceedings of Interspeech, Firenze, Italy*, pp. 861-864, 2011.

[54] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[55] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters,* vol. 49, pp. 1569-1570, 2013.

[56] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 5337-5341.

[57] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *arXiv preprint arXiv:1504.00923,* 2015.

[58] Y. Song, X. Hong, B. Jiang, R. Cui, I. McLoughlin, and L.-R. Dai, "Deep bottleneck network based i-vector representation for language identification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[59] L. F. Lamel and J.-L. Gauvain, "Cross-lingual experiments with phone recognition," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 1993, pp. 507-510.

[60] L. F. Lamel and J.-L. Gauvain, "Language identification using phone-based acoustic likelihoods," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, 1994, pp. I/293-I/296 vol. 1.

[61] K. M. Berkling, T. Arai, and E. Barnard, "Analysis of phoneme-based features for language identification," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference* 1994, pp. I/289-I/292 vol. 1.

[62] Y. Yan and E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, 1995, pp. 3511-3514.

[63] F. Pellegrino and R. André-Obrecht, "Automatic language identification: an alternative approach to phonetic modelling," *Signal Processing,* vol. 80, pp. 1231-1244, 2000.

[64] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 1988, pp. 501-541.

[65] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 271-284, 2007.

[66] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[67] R. W. Ng, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Analysis and selection of prosodic features for language identification," in *Asian Language Processing, 2009. IALP'09. International Conference on*, 2009, pp. 123-128.

[68] J. Foil, "Language identification using noisy speech," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, 1986, pp. 861-864.

[69] F. J. Goodman, A. F. Martin, and R. E. Wohlford, "Improved automatic language identification in noisy speech," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 international conference on*, 1989, pp. 528-531.

[70] R. W. Ng, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Prosodic attribute model for spoken language identification," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 5022-5025.

[71] A. Sangwan, M. Mehrabani, and J. H. Hansen, "Automatic language analysis and identification based on speech production knowledge," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 5006-5009.

[72] S. Mohanty and B. K. Swain, "Language identification using support vector machine," *Proceedings of OCOCOSDA-2010, Nepal*, 2010.

[73] J. Farinas and F. Pellegrino, "Automatic rhythm modeling for language identification," in *Seventh European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2001.

[74] J.-L. Rouas, J. Farinas, and F. Pellegrino, "Automatic modelling of rhythm and intonation for language identification," in *15th International Congress of Phonetic Sciences (15th ICPhS)*, 2003, pp. 567-570.

[75] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, pp. 436-456, 2005.

[76] T. Nagarajan and H. A. Murthy, "Language identification using parallel syllable-like unit recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, 2004, pp. I-401.

[77] D. Martinez, L. Burget, L. Ferrer, and N. Scheffer, "iVector-based prosodic system for language identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, Japan, 2012, pp. 4861-4864.

[78] S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-independent, text-independent language identification by HMM," in *ICSLP*, 1992, pp. 1011-1014.

[79] T. J. Hazen and V. W. Zue, "Segment-based automatic language identification," *The Journal of the Acoustical Society of America*, vol. 101, pp. 2323-2331, 1997.

[80] M. O. Eltayeb and M. E. Mustafa, "Acoustic-support vector machines approach to detect spoken Arabic language," in *Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on*, 2013, pp. 525-529.

[81] P. A. Torres-Carrasquillo, E. Singer, T. Gleason, A. McCree, D. A. Reynolds, F. Richardson, *et al.*, "The MITLL NIST LRE 2009 language recognition system," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4994-4997.

[82] L. J. Rodriguez-Fuentes, N. Brümmer, M. Penagarikano, A. Varona, G. Bordel, and M. Diez, "The albayzin 2012 language recognition evaluation," in *INTERSPEECH*, 2013, pp. 1497-1501.

[83] J. Gonzalez-Dominguez, I. Lopez-Moreno, P. J. Moreno, and J. Gonzalez-Rodriguez, "Frame-by-frame language identification in short utterances using deep neural networks," *Neural Networks,* vol. 64, pp. 49-58, 2015.

[84] A. Lozano-Diez, R. Zazo Candil, J. González Domínguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "An end-to-end approach to language identification in short utterances using convolutional neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[85] R. Zazo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks," *PloS one,* vol. 11, p. e0146917, 2016.

[86] W. Geng, W. Wang, Y. Zhao, X. Cai, and B. Xu, "End-to-End Language Identification Using Attention-Based Recurrent Neural Networks," in *INTERSPEECH*, 2016, pp. 2944-2948.

[87] B. Jiang, Y. Song, S. Wei, J.-H. Liu, I. V. McLoughlin, and L.-R. Dai, "Deep bottleneck features for spoken language identification," *PloS one,* vol. 9, p. e100795, 2014.

[88] P. Matejka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma*, et al.*, "Neural network bottleneck features for language identification."

[89] B. Ma, C. Guan, H. Li, and C.-H. Lee, "Multilingual speech recognition with language identification," in *INTERSPEECH*, 2002.