I.J. Information Technology and Computer Science, **2018, 8, 11-17**
Published Online August 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.08.02

Automatic Spoken Language Recognition with Neural Networks

Valentin Gazeau
Department of Computer Science at Sam Houston State University, Huntsville, TX, USA
E-mail: vcg006@shsu.edu

Cihan Varol
Department of Computer Science at Sam Houston State University, Huntsville, TX, USA
E-mail: cxv007@shsu.edu

Received: 22 June 2018; Accepted: 15 July 2018; Published: 08 August 2018

***Abstract*—**Translation has become very important in our generation as people with completely different cultures and languages are networked together through the Internet. Nowadays one can easily communicate with anyone in the world with the services of Google Translate and/or other translation applications. Humans can already recognize languages that they have priory been exposed to. Even though they might not be able to translate, they can have a good idea of what the spoken language is. This paper demonstrates how different Neural Network models can be trained to recognize different languages such as French, English, Spanish, and German. For the training dataset voice samples were choosed from Shtooka, VoxForge, and Youtube. For testing purposes, not only data from these websites, but also personally recorded voices were used. At the end, this research provides the accuracy and confidence level of multiple Neural Network architectures, Support Vector Machine and Hidden Markov Model, with the Hidden Markov Model yielding the best results reaching almost 70 percent accuracy for all languages.

***Index Terms*—**Hidden Markov Model, Language Identification, Language Translation, Neural Networks, Support Vector Machine.

I. Introduction

There are roughly 6,500 spoken languages available in the world today. While some of them are popular, i.e. Chinese spoken by over 1 Billion people, there are others such as Liki, Njerep which are only used by less than 1,000 people. Definitely, understanding the spoken language and corresponding accordingly when needed are vital to create communication between different background and language speaking people.

Spoken language recognition refers to the automatic process that determines the identity of the language spoken in a speech sample. This technology can be used for a wide range of multilingual speech processing applications, such as spoken language translation and/or multilingual speech recognition [1]. In practice, spoken language recognition is far more challenging than text-based language recognition because there is no guarantee that a machine is able to transcribe speech to text without errors. We know that humans recognize languages through a perceptual or psycho acoustic process that is inherent to the auditory system. Therefore, the type of speech perception that human listeners use is always the source of inspiration for automatic spoken language recognition [2].

This paper outlines how Neural Networks can be trained via Support Vector Machine and Hidden Markov Model to automatically identify the language directly from speech using libraries such as TensorFlow [3]. This can be done in many different ways as the results depend on a number of factors: for example the diversity and size of the training data, and/or the Neural Network model. It also demonstrates how the training data was gathered, the problems faced and how testing was conducted.

The paper is organized as follows: Section II introduces previous attempts at automatic spoken language identification using phoneme classifiers, statistical recurrent Neural Networks, and deep learning approaches. Section III covers what models were chosen to implement the Neural Network as well as how the training data was gathered and how the training was conducted. Section IV details how the testing was performed and how the performance of the multiple models were evaluated as well as the results. The conclusion is drawn out in Section V along with the future improvements that can be made. Your goal is to simulate the usual appearance of papers in a Journal of the Academy Publisher. We are requesting that you follow these guidelines as closely as possible.

II. Related Works

Several attempts have already been made to recognize spoken languages with different data sets. While Neural Networks was used for variety of identification/classification problems [4, 5], this section covers some of the most promising recent works in automatic spoken language identification using Neural Networks: Srivastava et al. [6] use a language independent phoneme classifier that extracts the sequence of phonemes from an audio file. The phoneme sequence is then classified using statistical and recurrent Neural Network models (RNNs). With phoneme classification of three languages (Turkish, Uzbek and Mandarin) authors achieved an average accuracy of 58%.

Montavon's [7] attempt uses a deep neural network featuring three convolutional layers as well as a smaller attempt with a single convolutional layer time-delay model. Trying to classify English, German, and French, author achieves 91.3% accuracy for known speakers (used in training) and 80% for unknown speakers.

Lei et. al proposed two novel frontends for robust language identification (LID) using a convolutional neural network (CNN) trained for automatic speech recognition (ASR) [8]. In the CNN/i-vector frontend, the CNN is used for getting the posterior probabilities for i-vector training and extraction. The authors evaluated their approach on heavily low quality speech data, but still they were able to achieve significant improvements of up to 50% on average equal error rate compared to a universal background model.

Another trial was done by Lee et al. [9] through the use of unsupervised convolutional deep belief networks to learn the phonemes from speech data. They used spectrograms of 20ms with 10ms overlaps from the English-only corpus. As another research, they used unsupervised convolutional deep belief networks for gender identification in audio files. Graves et al. used recurrent neural network for language identification, they found that a deep long short-term memory recurrent neural network can achieve a test set error of 17.7% on the TIMIT phoneme recognition benchmark [10].

House and Neuberg [11] worked on phonetics data and made useful information speech instead of acoustic features using secret Markov Model (HMM) for training of system in eight languages with 80% accuracy. Foil used two different approaches to learn the language with noisy background. The first approach processed pitch and energy contours and then language features for language definition. Calculation in the second approach formant vectors (K-domain clustering algorithm) are made to distinguish the same phonemes for different languages [12].

Song et al. presented a unified i-vector framework for language identification based on deep bottleneck networks trained for automatic speech recognition. The output from different layers of a deep bottleneck networks are exploited to improve the effectiveness of the i-vector representation Authors evaluated their approach on Arabic dialect recognition and achieved higher accuracy compared to original deep bottleneck networks [13].

Jothilakshmi et. al. developed a two level language identification system for Indian languages using acoustic features. First, the system identifies the family of the spoken language, second, categorize the language in the corresponding family. The system is modeled using hidden Markov model (HMM), Gaussian mixture model (GMM) and artificial neural networks (ANN). The authors shown that GMM is performing well with 80.56% accuracy [14].

Several approaches have been proposed in the literature that use support vector machines for speech applications. The first approach tried to model emission probabilities for hidden Markov models [15, 16]. This approach has been moderately successful in reducing error rates, but it suffers from several problems. First, large training sets result in longer training times for support vector methods. Second, since there is no possibility of the support vector machine coming out, the emission probabilities should be approximated [17]. This approach is necessary to combine probabilities using the standard framework independence method used for loudspeaker and language recognition. The second set of approaches tries to combine Gaussian Mixture Model approaches with SVMs [18]. A third set of approaches are based upon comparing sequences using the Fisher kernel proposed by Jaakkola and Haussler [21].

Although most of these neural networks have a relatively good accuracy for language identification, variety of test and training data is limited to give an idea about how those data affect the performance of language identification [22]. Moreover, to the best of our knowledge, this would be the only study to focus on French, English, Spanish, and German languages altogether which are categorized as widely used languages in the USA [21].

As discussed above, there are plenty of different neural network models that can be used to achieve spoken language identification. However, there are two main approaches for speech processing: static classification, which involves classifying based on the entirety of the speech file, and segmentation. Segmentation consists of splitting the speech recording in smaller fixed-length segments that are individually classified, each classification is then recorded and the class that has the most corresponding segments is chosen as the predominant class for the whole speech file. In this work the two classification techniques are going to be tested and contrasted in order to achieve maximum accuracy and confidence level. For static classification the Support Vector Machine neural network model will be used, while the Hidden Markov Model architecture will be used for segmented classification.

III. Methodology

A. Model: Support Vector Machine

The support vector machine (SVM) is a supervised learning model developed in the 1990s with associated learning algorithms that analyze data used for classification and regression [22]. In the case of SVM, it learns a linear model, finding the optimal hyperplane, for linearly separable patterns [23]. It also extends to patterns that are not linearly separable using transformations of original data to map into the new space.

Assume we are given a data set of n elements in the form: $(\dot{x}_1, y_1), ..., (\dot{x}_n, y_n)$ where y_i corresponds to a specific class from which $\dot{x}_i$ belongs. Then each x_i is a multi-dimensional real vector, the SVM will find the maximum-margin hyperplane that divides the groups of $\dot{x}_i$ points that correspond to a specific y_i class from the other classes. The hyperplane is found with: $\dot{w}.\dot{x}$ - b = 0 where w is the normal vector to the hyperplane (Fig 1).

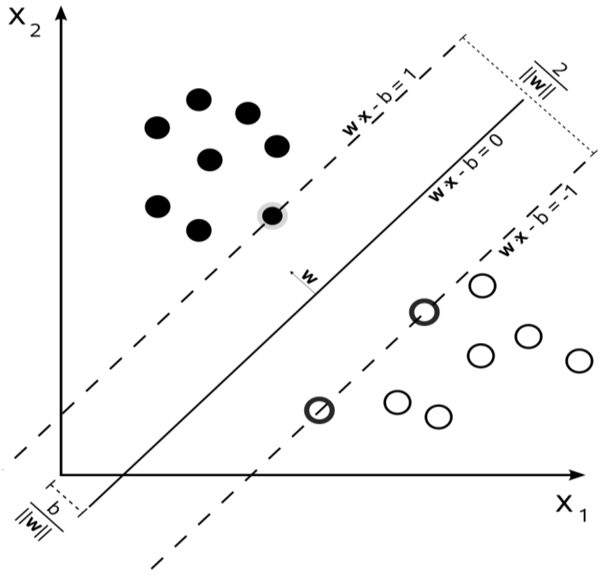

Fig.1. Support Vector Machine

B. Model: Hidden Markov Model

Hidden Markov Models (HMMs) provide a simple and effective framework for modeling time-varying spectral vector sequences, which explains why most present day large vocabulary continuous speech recognition systems are based on this model [24]. Since speech has temporal structure and can be encoded as a sequence of spectral vectors spanning the audio frequency range, the hidden Markov model (HMM) provides natural frameworks for constructing such models

Time-varying spectral vector sequences can be depicted as the following: $x(t)$ Ɛ $(x1, x2, x3, . . .)$ with $x(t)$ as random variables at a given time t. HMM neural networks can easily represent each element as a random variable $x(t)$ in a hidden state at time t . The random variable $y(t)$ is the observation at time t with its own set $y(t)$ Ɛ $(y1, y2, y3, . . .)$ depicted in Fig. 2.

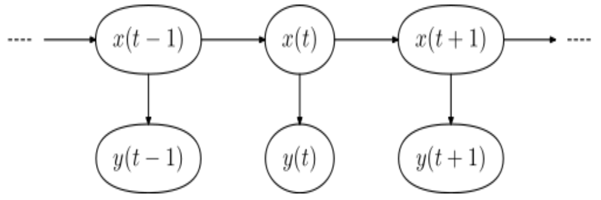

Fig.2. Hidden Markov Model

C. Segment Classification

One of the first improvements that was made to the original Neural Network design using PyAudioAnalysis [25] is to classify a sound not based on the entirety of the speech file but rather on smaller segments of the file. This allows to leave out the ”blank” or ”noise-free” part of every recording and only analyze speech as shown in Fig. 3. This way it will actually classify every 0.5 seconds of the speech recording to analyze to produce statistics that are redirected to an output file. The file can then simply be read and interpreted with the data indicating the probability of classification (confidence) for a certain language based on the classification of each segment in that speech recording.

This is especially helpful for longer speech files and also help when different languages are spoken within the same speech file, the Neural Network will be able to accurately determine which segments correspond to which language and choose the most predominant language in order to apply speech recognition and translation approach maintains higher recall rate and precision in the voluminous retrieval context.

D. Data Acquisition and Issues

During the first attempt, the main issue while training was actually the lack of data. The data was downloaded from shtooka.net which has plenty of audio files. The problem came from the fact that most recordings are done by a single person. This did not help the Neural Network because without diversity it will try to extract other features such as voice and/or accent, making more of a Speaker Identification Neural Network rather than Language Identification Network.

Not being able to find reliable sources that had easy access to massive amounts of audio recordings made it a challenge. The four languages (French, English, Spanish, and German) needed to be diverse with multiple different speaker, the use of python scripts to automatically download them from the following sources helped the process.

VoxForge [26] is an open-source speech audio corpus containing samples from more than 18 different languages. Data consists of short user uploaded audio files and a machine transcription of the spoken text. All samples are about 5-10 seconds long and have varying speakers totaling a duration of about 5 hours of speech per language (about 5GB). It is important to note that the quality of different samples varied based on the recording equipment used by the speaker.

YouTube [27] seemed like the best place to download large amounts of audio-only files from different language-speaking channels. Videos from popular news channels such as CNN and France24 which contained hundreds of videos were downloaded. One of the main issues with the YouTube data set is that it contains irregularity in speech length and/or speaker for most of its elements. Moreover, in contrast to the artificial nature of VoxForge, the YouTube data set usually has several persons speaking to each other. The resulting speech sounds more natural but may be faster paced in general. The recording quality, on the other hand, is very high and noise free. Even though most of the content has no background noise, news stories will occasionally feature clips that do not contain any speech data.

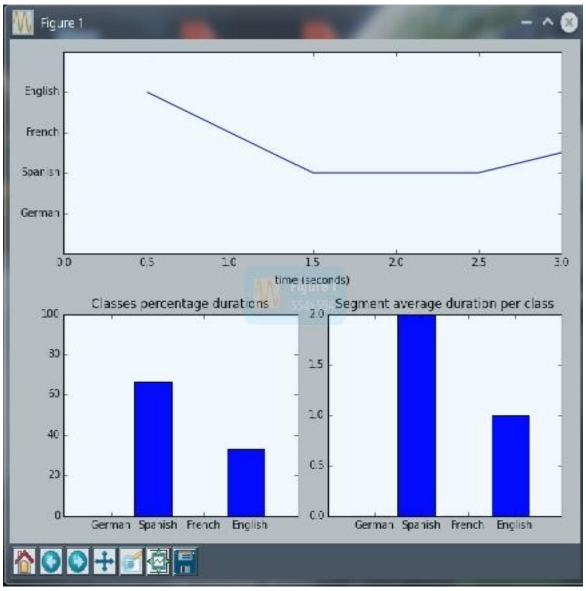

Fig.3. Speech Segmentation

E. Training

Once the data has been gathered, the Neural Network was trained. The first few attempts using the data gathered from Shtooka.com only had about 2,000 recordings per language. The training took about 12 hours with 4,000 total speech files, on an Intel Celeron N2940 processor (2MB Cache, Up To 2.25GHz) and 2GB of RAM. Now with the data gathered from YouTube, about 1000 (500 from each channel) recordings of variable length resulting in about 5GB of speech files per languages for a total of 20GB of added training data. This took approximately 7 days of training time. Likewise, for the data that was gathered from VoxForge, 2,000 variable-length files about 23,000 of short speech files from a wide range of people were available for training. This resulted in about 20GB of data as well. The training took about a week also.

IV. Performance

This section covers how testing was conducted to evaluate the performance of each Neural Network. Before training, 20 randomly selected audio files from each language's training directory (eg: 20 for French, etc...) were removed. These files are used as sample data to evaluate the Neural Network. Testing was performed every time when there was a change to the Neural Network, whether it was a change of training data and/or change of Neural Network architecture. Multiple attempts were necessary to try out different Neural Network architectures such as: HMM (Hidden Markov Model) and SVM (Support Vector Machines). Table 1 shows a list of all attempts that were conducted using the two architectures with three different data sets. The SVM model was only used once since after testing the two models with the same training data sets, HMM performed relatively better as it is a more suited model for time-varying spectral vector such as speech.

A. Testing

When testing the Neural Networks, for the first seven test cases, 20 randomly selected audio files were removed from each language's training directory. These files are used as sample data to evaluate the Neural Network. Another testing phase using our own voice is then performed for comparison using 10 preset sentences for the same seven test cases. The evaluation is measured by the accuracy of the Neural Network: the percentage of certitude for the classification of a sample speech file. The 10 preset sentences are inputted via the Recorder that was implemented alongside the Neural Network.

Table 1. Neural Network Architectures

ID	Model	Dataset	Classified Language
1	HMM	Shtooka	French, English
2	SVM	Shtooka	French, English
3	HMM	Shtooka	French, English, Spanish, German
4	HMM	VoxForge	French, English
5	HMM	VoxForge	French, English, Spanish, German
6	HMM	Youtube	French, English
7	HMM	Youtube	French, English, Spanish, German
8	HMM	VoxForge –	French, English, Spanish, German

Based on the performances shown after these initial tests, the test data was increased to reflect the effectiveness of the Neural Network, test case #8. The details of this test will be discussed in the 4.2 Results section the paper. Neither of the speech recognition or translation had to be tested as they were entirely implemented by Google and work as intended. The only kind of errors encountered with them were that when the language classification failed and it tried to recognize speech in the wrong language. For example if ”Hi, how are you?” is given as input and the Neural Network would classify the language as French, it would attempt to apply the French setting for speech recognition which usually gives something completely different.

B. Results

Tables 2 & 3 reflects the performances of the Neural Networks recognizing only French and English followed by the table with the performances for all four target languages respectively (French, English, Spanish, and German). The results show first the testing involving the saved 20 sentences from each language (Sample Data - SD) and also the testing of the 10 sentences inputted via microphone by ourselves (Recorder - REC). These tables showcase the accuracy which is simply the percentage of sentences where the language was guessed correctly and the average confidence which is an average of the language probability for each sentences (degree of certainty). We can clearly see that the training data affects both accuracy and confidence of the Neural Network with the diversity of the training data from VoxForge yielding the best results.

C. Interpretation

The first two attempts used the same training data set, but first attempt used a Hidden Markov Model and the second attempt used the SVM model. We can see from Table II that the Hidden Markov Model performed much better than the Support Vector Machine. The SVM model was dropped and HMM was accepted as the Neural Network of choice for Automatic Language Identification for later tests.

Since the first two attempts were not very accurate, lack of class types were believed to be the reason. Adding more languages to classify would help the problem. But the results of third attempt shows that it wasn't the case. The real reason was that the Neural Network was trained to recognize the speaker, the voice used for recording resembled the voice used in the English data set, explaining why for most attempts it classified the recording as English hence the low accuracy scores.

The sixth and seventh attempts using HMMs and the data gathered from the YouTube script didn't yield very good results and there are multiple explanations for that: first, the data was collected from two different news channels for each languages meaning that there isn't that much diversity, it will always be the same people talking with the same voices and accents. Another is that the recordings are very long so instead of having many short speech files to train the network, it was trained for the entirety of the speech file that can sometimes be more than 15 minutes. Another reason is the fact that some of these videos included moments without speech or moments with other sounds, which will be attributed to the target language when trained.

The fourth and fifth attempts were the most successful as the training data was much more vast and rich. That is instead of training it with 2,000 words from the same speaker for each language, the Neural Network was trained with the data acquired with the scripts from VoxForge. The data comprised about 23,000 words per languages, including thousands of different speakers with different voices, accents, and recording equipment. This forced the Neural Network not to recognize the voice, accent, or background noise of a particular speech file but rather the corpus of the language itself. The fifth attempt shows less accuracy than the fourth attempt that makes sense since there are more languages to classify, increasing its rate of failure.

Table 2. Accuracy and Confidence Level for Two Language Classification

ID and Dataset	Accuracy of French	Confidence of French	Accuracy of English	Confidence of English
1 (SD)	75%	88.32%	80%	91.57%
2 (SD)	60%	83.65%	70%	89.70%
4 (SD)	95%	95.22%	90%	93.97%
6 (SD)	75%	41.45%	70%	44.17%
1 (REC)	50%	66.42%	60%	73.09%
2 (REC)	40%	50.76%	50%	61.21%
4 (REC)	80%	83.82%	80%	85.13%
6 (REC)	60%	21.25%	50%	33.64%

Table 3. Accuracy and Confidence Level for Four Language Classification

ID and Dataset	Accuracy of French	Confidence of French	Accuracy of English	Confidence of English
3 (SD)	55%	64.25%	70%	85.09%
5 (SD)	85%	90.53%	95%	97.38%
7 (SD)	40%	21.10%	35%	19.70%
3 (REC)	40%	33.62%	50%	47.23%
5 (REC)	60%	65.35%	80%	77.44%
7 (REC)	30%	26.44%	20%	20.34%
ID and Dataset	Accuracy of Spanish	Confidence of Spanish	Accuracy of German	Confidence of German
3 (SD)	60%	69.07%	65%	73.50%
5 (SD)	85%	91.17%	90%	95.86%
7 (SD)	40%	20.21%	50%	38.97%
3 (REC)	30%	28.86%	40%	35.57%
5 (REC)	60%	66.83%	70%	76.96%
7 (REC)	30%	26.07%	40%	34.22%

Since the Neural Network trained with the VoxForge data set performed the best, some more testing was done with a greater range of sample data in order to have a more precise accuracy rate (Table 4). The last attempt used the same Neural Network as attempt #5 but instead of testing it with only the 20 sample data for each language, this time it is tested with 500 speech recordings per language for 2,000 speech recordings (10% of data set). These sample data files were collected separately from VoxForge meaning that they count as unknown speakers since they were not used for training. A python script was used to automate the testing process, redirecting results to a text file.

Table 4 shows the results. The accuracy for each language converge to an average of 66.55% which makes sense considering that the more classes the Neural Network has to memorize, the less precise the classification process will be. With a richer training data set the Neural Network can surely achieve higher than 70% accuracy since an accurate and diverse training data is the most important aspect of a neural network.

Table 4. Accuracy and Confidence Level for Four Language Classification with Larger Dataset

ID and Dataset	Accuracy of French	Confidence of French	Accuracy of English	Confidence of English
8 (SD)	64.80%	71.12%	67.40%	65.43%
	Accuracy of Spanish	Confidence of Spanish	Accuracy of German	Confidence of German
	69.80%	73.20%	64.20%	67.25%

V. Conclusion and Future Work

For automatic language identification, arguably, the most important aspect is to gather as much data as possible. In order to improve both the accuracy and average confidence, it is necessary for the data to be rich in diversity because of the risk of training the Neural Network to classify accents, genders, voices etc... instead of classifying the language as intended. Therefore, by training it with a more diverse and richer data set including both genders, all ages, and all accents, the Neural Network has to extract information from the features of the language.

There are multiple aspects of this study can be improved. One of them is the architecture, it was obvious that the Hidden Markov Model was a better fit in this case than the Support Vector Machine but that does not mean HMM is the best for speech audio analysis. Especially if an immense data set is available for training the Neural Networks. In that case, the best architecture would most certainly be one of the deep learning alternative. Deep learning models are loosely related to information processing and communication patterns in a biological nervous system, such as neural coding that attempts to define a relationship between various stimuli and associated neuronal responses in the brain.

Deep learning architectures such as deep Neural Networks, deep belief networks and recurrent Neural Networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bio informatics and drug design, where they have produced results comparable to and in some cases superior to human experts.

References

[1] I. Lopez-Moreno et al, "Automatic Language Identification Using Deep Neural Networks," *2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. Florence, Italy. 2014, 4(9).

[2] D. Cazzani, "Audio Processing in TensorFlow," Retrieved from: https://towardsdatascience.com/audio-processing-in-tensorflow-208f1a4103aa. 2017, 6(30).

[3] TensorFlow, Retrieved from: https://www.tensorflow.org/. 2018, 5(12).

[4] D. Pawade, A. Sakhapara, M. Jain, N. Jain and K. Gada, "Story Scrambler – Automatic Text Generation Using Word Level RNN-LSTM", *International Journal of Information Technology and Computer Science (IJITCS),* Vol.10, No.6, pp.44-53, 2018. DOI: 10.5815/ijitcs.2018.06.05

[5] M. Sreeshakthy, J. Preethi and A. Dhilipan, "A Survey on Emotion Classification From Eeg Signal Using Various Techniques and Performance Analysis", *International Journal of Information Technology and Computer Science(IJITCS),* Vol.8, No.12, pp.19-26, 2016. DOI: 10.5815/ijitcs.2016.12.03

[6] B. M. L. Srivastava, H. K. Vydana A. K. Vuppala and M. Shrivastava, "A Language Model Based Approach Towards Large Scale and LightWeight Language Identification Systems," *arXiv preprint arXiv:1510.03602*. 2015, 10.

[7] G. Montavon, "Deep Learning for Spoken Language Identification," *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*. 2009

[8] Y. Lei, L. Ferrer, A. Lawson, M. McLaren and N. Scheffer, "Application of Convolutional Neural Networks to Language Identification in Noisy Conditions," *The Speaker and Language Recognition Workshop*. Joennsu, Finland. 2014, 6(16-19), pp. 287-292.

[9] H. Lee, Y. Largman, P. Pham and A. Y. Ng, "Unsupervised Feature Learning for Audio Classification using Convolutional Deep Belief Networks," *NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems Pages*. Vancouver, British Columbia, Canada. 2009, 11 (7-10), pp. 1096-1104.

[10] A. Graves, A. Mohamed and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *2013 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Vancouver, British Columbia, Canada. 2013, 5 (26-31).

[11] A. S. House and E. P. Neuberg, "Toward Automatic Identification of Language of an Utterance. I. Preliminary Methodological Considerations," *Journal of the Acoustical Society of America*, 1977, (62,3,), pp. 708-713.

[12] J. T. Foil, "Language identification using noisy speech," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1986, pp. 861–864.

[13] Y. Song et al, "Deep bottleneck network based i-vector representation for language identification," *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association.* Dresden, Germany. 2015, 9 (6-10), pp. 398-402.

[14] S. Jothilakshmi, V. Ramalingam and S. Palanivel, "A Hierarchical Language Identification System for Indian Languages," *Digital Signal Processing*, 2012. 22(3), pp. 544-553

[15] V. Wan and W. M. Campbell, "Support vector machines for verification and identification, in: Neural Networks for Signal Processing X," *Proceedings of the 2000 IEEE Signal Processing Workshop*, 2000, pp. 775–784.

[16] A. Ganapathiraju and J. Picone, "Hybrid SVM/HMM architectures for speech recognition," *Speech Transcription Workshop*. 2000.

[17] J. C. Platt, "Probabilities for SV machines" A.J. Smola, P.L. Bartlett, B. Schölkopf, D. Schuurmans (E ds.), *Advances in Large Margin Classifiers,* MIT Press, Cambridge, MA 2000, pp. 61-74

[18] J. Kharroubi, D. Petrovska-Delacretaz and G. Chollet, "Combining GMMs with support vector machines for text-independent speaker verification," *Eurospeech*. 2001, pp. 1757–1760

[19] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers" M.S. Kearns, S.A. Solla, D.A. Cohn (Eds.), *Advances in Neural Information Processing,* vol. 11, MIT Press, Cambridge, MA 1998, pp. 487-493

[20] Y. Kumar and N. Singh, "Automatic Spontaneous Speech Recognition for Punjabi Language Interview Speech Corpus", *International Journal of Education and Management Engineering (IJEME),* Vol.6, No.6, pp.64-73, 2016.DOI: 10.5815/ijeme.2016.06.07

[21] J. Burton, "The Most Spoken Languages in America," Retrieved from: https://www.worldatlas.com/articles/the-most-spoken-languages-in-america.html. 2017, 4(25).

[22] J. Weston, "Support Vector Machine Tutorial," Retrieved from: ftp://ftp.umiacs.umd.edu/pub/chenxi/Project%20FTP/Finder_FTP/svmlib/jason_svm_tutorial.pdf . 2014

[23] P. Thamilselvana and J. G. R. Sathiaseelan, "A Comparative Study of Data Mining Algorithms for Image Classification," *International Journal of Education and Management Engineering (IJEME),* Vol.2, No.9, pp.1-9, 2015.DOI: 10.5815/ijeme.2015.02.01

[24] K. Tumilaar, Y. Langi and A. Rindengan, "Hidden Markov Model," *De Cartesian*, 2015, 4(1)

[25] Pyaudioanalysis 0.1.3. "Python Package Index," Retrieved from: https://pypi.python.org/pypi/pyAudioAnalysis/. 2018.

[26] Voxforge Free Speech Recognition (Linux, Windows and Mac). Retrieved from: http://www.voxforge.org/. 2018.

[27] Youtube. Retrieved from: https://www.youtube.com/. 2018.

Authors' Profiles

Valentin Gazeau: PhD student for Digital and Cyber Forensics program at Sam Houston State University. His major research areas are programming languages, neural network, and language and speaker identification.

Cihan Varol: Associate Professor and Graduate Coordinator for the Department of Computer Science and Sam Houston State University. His research interests are in the general area of information (data) quality, VoIP Forensics, and risk management with specific emphasis on personal identity recognition, record linkage, entity resolution, pattern matching techniques, natural language processing, multi-platform VoIP applications, VoIP artifacts data cleansing, and quality of service in business process automation. These studies have led to more than 50 peer-reviewed journal and conference publications, and two book chapters.

How to cite this paper: Valentin Gazeau, Cihan Varol, "Automatic Spoken Language Recognition with Neural Networks", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.8, pp.11-17, 2018. DOI: 10.5815/ijitcs.2018.08.02