IADIS International Journal on Computer Science and Information Systems
Vol. 15, No. 1, pp. 99-113
ISSN: 1646-3692

IMPROVED VOICE-BASED BIOMETRICS USING MULTI-CHANNEL TRANSFER LEARNING

Youssouf Ismail Cherifi[1] and Abdelhakim Dahimene[1,2]
[1]*Universite de M'Hamed Bougara, Boumerdes, Algeria*
[2]*Signal and System laboratory, Boumerdes, Algeria*

ABSTRACT

Identifying the speaker has become more of an imperative thing to do in the modern age. Especially since most personal and professional appliances rely on voice commands or speech in general terms to operate. These systems need to discern the identity of the speaker rather than just the words that have been said to be both smart and safe. Especially if we consider the numerous advanced methods that have been developed to generate fake speech segments. The objective of this paper is to improve upon the existing voice-based biometrics to keep up with these synthesizers.

The proposed method focuses on defining a novel and more speaker adapted features by implying artificial neural networks and transfer learning. The approach uses pre-trained networks to define a mapping from two complementary acoustic features to a speaker adapted phonetic features. The complementary acoustics features are paired to provide both information about how the speech segments are perceived (type 1 feature) and produced (type 2 feature). The approach was evaluated using both a small and large closed-speaker data set. Primary results are encouraging and confirm the usefulness of such an approach to extract speaker adapted features whether for classical machine learning algorithms or advanced neural structures such as LSTM or CNN.

KEYWORDS

Speech Analysis, Transfer Learning, Pattern Recognition, Speaker Recognition, Feature Extraction

1. INTRODUCTION

Modern age devices have become more user friendly and natural to interact with than ever, and perhaps there is no functionality that made this true then voice-user interface (VUI) (Syntellect Inc. 2003). VUI uses speech recognition technology to make spoken human interaction with machines a possibility. VUI has been added to a variety of devices such as automobiles, home automation systems, computer operating systems, phones, and even home appliances like washing machines (Donald 1983), microwaves, and televisions have all become voice commanded devices VCD. In fact, as of 2019, an estimated 3.25 billion VCD were used across the world and by 2023 this number will reach approximately eight billion units (Statista 2020). Given the wide application of VUI, it has become imperative to identify not just the words that have been spoken but also who is issuing these commands to make modern VCD both smart and safe. However, implementing a voice-based biometric system is not an easy task given the various Deepfake algorithms that can generate compelling speech segments by having access just to 5 seconds of the target's voice (Ye et al., 2019).

The objective of this paper is to improve existing voice-based biometric systems with the hope of making them more robust and immune to deepfakes. To that end, the general structure of voice-based biometrics systems is considered (Figure 1).

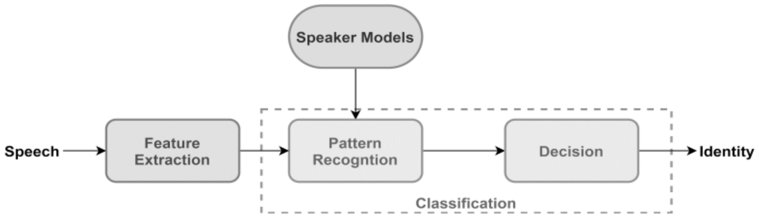


Figure 1. The general structure of speaker recognition systems

Most of the research that has been done in order to improve the accuracy of speaker recognition tasks focuses on adapting high-performing pattern matching algorithms to speaker recognition. For instance algorithms such as Gaussian Mixture Model (GMM) (Chakroum et al, 2016), Support Vector Machine (SVM) (Campbell et al., 2006), and Artificial Neural Network (ANN) (Srinivas et al, 2014) have all been suggested to obtain a better recognition rate. Continuous research and effort are ongoing, involving the combination of two modeling techniques (Al-Shayea & Al-Ani, 2016. Chakroborty & Saha, 2009. Singh et al, 2016. Awais et al, 2014) or the implementation of specific hardware (Gaafar et al, 2014).

Obtaining a better recognition rate does not depend only on the pattern matching block. In fact, the role of the feature extraction block is also important for recognition. Feature extraction when carried out correctly will ensure that the information used for matching fulfill the following criteria (Nolan, 1983; Wolf, 1972): i) easy to extract ii) high inter-speakers variability and intra-speaker consistency iii) difficult to mimic/impersonate vi) unaffected by health or age. There are several features to choose from when it comes to feature extraction. However, the most prominent features are acoustic features. These features capture the spectral parameters of the speech signal. This for example could be the psychoacoustic (i.e how sound is being perceived) information provided by the so-called Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980). MFCCs provides better accuracy compared to other features (Kinnunen et al., 2007; Thian et al., 2004). Alternative MFCC driven features have been developed to emphasize speaker-specificity (Charbuillet et al., 2006; Miyajima et al., 2001; Kinnunen, 2002; Orman and Arslan, 2001). Alternative information represented by acoustic features is how sound is produced. This information is estimated by linear predictors (LP) (Makhoul, 1975; Mammone et al., 1996). LP coefficients are not very powerful as a speaker-specific feature. That is why they have been improved and adapted to a more robust and less correlated feature such as the linear predictive cepstral coefficients (LPCCs) (Huang et al., 2001), the perceptual linear prediction (PLP) coefficient (Hermansky, 1990). Acoustics features have been also combined with other feature levels such as phonologic (Murty and Yegnanarayana, 2006; Zheng et al., 2007) and semantic by providing more discriminatory tokens (Andrews et al., 2002; Campbell et al., 2004).

Based on the aforementioned work, an improved speaker recognition approach is suggested in this paper. The approach is to develop speaker-specific short-term features using transfer learning and to couple that with the use of classifiers (such as LSTM and CNN) that are able to incorporate the semantic information present in the speech in its entirety.

2. MATERIALS AND METHODS

2.1 Proposed Approach

The methodological backbone of this paper stems from the fact that for decades now we have been using the same kinds of features for both speech recognition and speaker recognition (Kinnunen and Li, 2010). The idea is to propose a speaker-specific mapping from prominent existing features by extending the finding of (Sinno and Qiang, 2010) to do so. For each frame, two parallel feature streams are extracted: a feature set representing the psychoacoustic information (e.g MFCC), and a feature set representing the physical parameters of the individual's speech system (e.g LPC, LPCC, or PLP). The two streams are used to find a mapping from the acoustic spectrum to the speaker-specific spectrum using the ANN shown below (Figure 2).

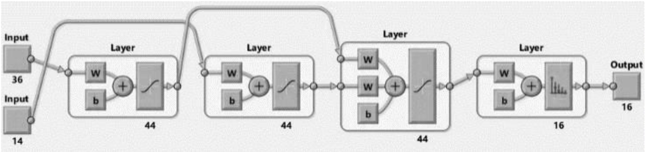
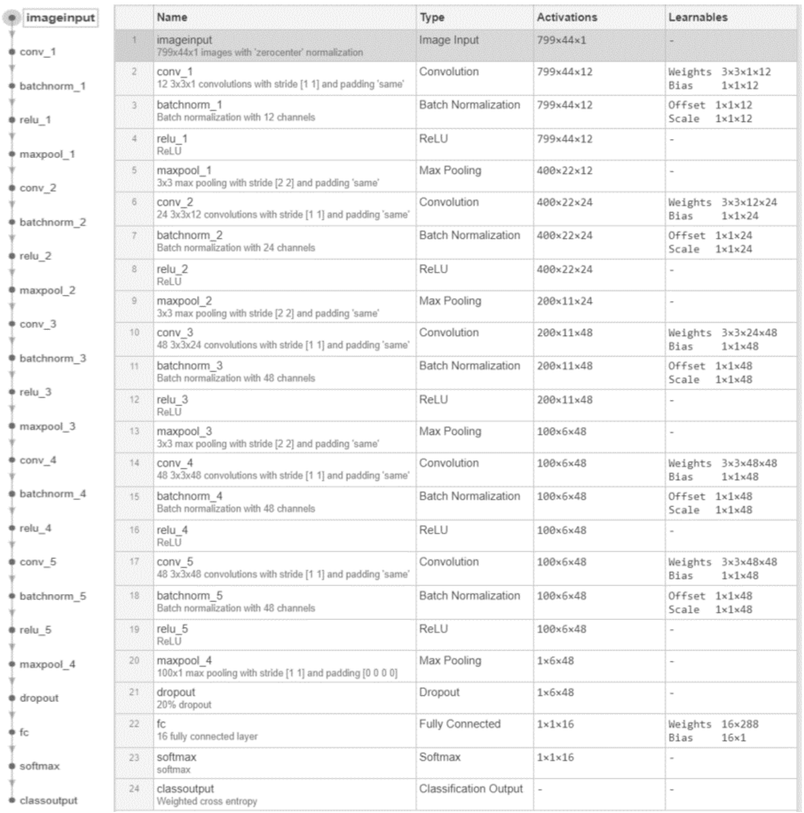


Figure 2. The proposed ANN structure for feature fusion. Top input is for type 1 feature and bottom input is for type 2

The structure shown above was implemented using MATLAB 2017a Deep Learning Toolbox (Figure 2). Each of the input layers as well as the hidden layer contained 44 perceptrons, to account for all possible phonemes in the English language. To reduce the influence of extreme values or outliers in the dataset without having to remove them, a SoftMax function was used as the activation function for the output layer. For input layers and hidden layers perceptrons, the tangential sigmoid activation function was used. This function has a steeper derivative which makes it a good candidate for extracting intermediary features (Meena et al, 2011).

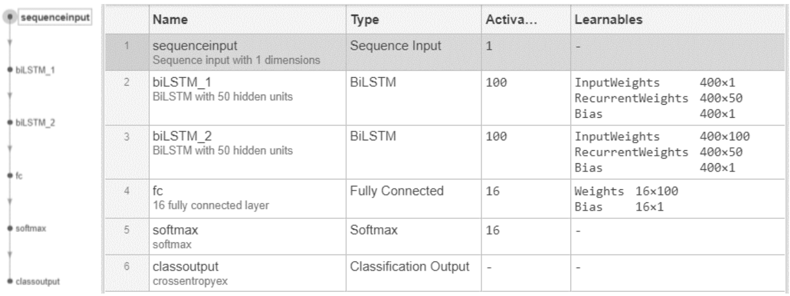
To train the structures shown above, the conjugate gradient backpropagation algorithm is used to reduce the Sum of Squared Errors (SSE) between the outputs of the network and a vector of desired targets. This algorithm has better accuracy when compared with other algorithms (Vacic, 2015).

The structure shown above has the ability to provide speaker-specific phonetic information from each frame by mapping the two streams of features into a single feature. The obtained stream can be viewed as a sequence of token that in addition to being specific speakers, can also contain a co-occurring pattern of pronouncing certain words that should emphasize the speaker's print. To that end, 2 additional structures were designed as tokenizers for the entire stream of the newly extracted feature. Both structures were implemented using a more recent version of MATLAB (R2020a).



	Name	Type	Activations	Learnables
1	imageinput 799x44x1 images with 'zerocenter' normalization	Image Input	799×44×1	-
2	conv_1 12 3x3x1 convolutions with stride [1 1] and padding 'same'	Convolution	799×44×12	Weights 3×3×1×12 Bias 1×1×12
3	batchnorm_1 Batch normalization with 12 channels	Batch Normalization	799×44×12	Offset 1×1×12 Scale 1×1×12
4	relu_1 ReLU	ReLU	799×44×12	-
5	maxpool_1 3x3 max pooling with stride [2 2] and padding 'same'	Max Pooling	400×22×12	-
6	conv_2 24 3x3x12 convolutions with stride [1 1] and padding 'same'	Convolution	400×22×24	Weights 3×3×12×24 Bias 1×1×24
7	batchnorm_2 Batch normalization with 24 channels	Batch Normalization	400×22×24	Offset 1×1×24 Scale 1×1×24
8	relu_2 ReLU	ReLU	400×22×24	-
9	maxpool_2 3x3 max pooling with stride [2 2] and padding 'same'	Max Pooling	200×11×24	-
10	conv_3 48 3x3x24 convolutions with stride [1 1] and padding 'same'	Convolution	200×11×48	Weights 3×3×24×48 Bias 1×1×48
11	batchnorm_3 Batch normalization with 48 channels	Batch Normalization	200×11×48	Offset 1×1×48 Scale 1×1×48
12	relu_3 ReLU	ReLU	200×11×48	-
13	maxpool_3 3x3 max pooling with stride [2 2] and padding 'same'	Max Pooling	100×6×48	-
14	conv_4 48 3x3x48 convolutions with stride [1 1] and padding 'same'	Convolution	100×6×48	Weights 3×3×48×48 Bias 1×1×48
15	batchnorm_4 Batch normalization with 48 channels	Batch Normalization	100×6×48	Offset 1×1×48 Scale 1×1×48
16	relu_4 ReLU	ReLU	100×6×48	-
17	conv_5 48 3x3x48 convolutions with stride [1 1] and padding 'same'	Convolution	100×6×48	Weights 3×3×48×48 Bias 1×1×48
18	batchnorm_5 Batch normalization with 48 channels	Batch Normalization	100×6×48	Offset 1×1×48 Scale 1×1×48
19	relu_5 ReLU	ReLU	100×6×48	-
20	maxpool_4 100x1 max pooling with stride [1 1] and padding [0 0 0 0]	Max Pooling	1×6×48	-
21	dropout 20% dropout	Dropout	1×6×48	-
22	fc 16 fully connected layer	Fully Connected	1×1×16	Weights 16×288 Bias 16×1
23	softmax softmax	Softmax	1×1×16	-
24	classoutput Weighted cross entropy	Classification Output	-	-

Figure 3. CNN based architecture for feature tokenization



	Name	Type	Activa...	Learnables
1	sequenceinput Sequence input with 1 dimensions	Sequence Input	1	-
2	biLSTM_1 BiLSTM with 50 hidden units	BiLSTM	100	InputWeights 400×1 RecurrentWeights 400×50 Bias 400×1
3	biLSTM_2 BiLSTM with 50 hidden units	BiLSTM	100	InputWeights 400×100 RecurrentWeights 400×50 Bias 400×1
4	fc 16 fully connected layer	Fully Connected	16	Weights 16×100 Bias 16×1
5	softmax softmax	Softmax	16	-
6	classoutput crossentropyex	Classification Output	-	-

Figure 4. BiLSTM based architecture for feature tokenization

The hyperparameters for each structure are as shown above (Figures 3, 4) and they were both trained using categorical class entropy as a loss function. This was reduced across training epochs using the Adam optimizer, which as better suited for this kind of complexity involving large data and parameters.

Overall, 3 sets of features were implied in this paper to provide the different pairs of parallel streams. The extraction was conducted also in MATLAB using the Auditory Toolbox (Slaney, 1998) and VOICEBOX (Brooks, 1997).

2.1.1 Mel Frequency Cepstral Coefficient

In practice, MFCCs are near impossible to beat as even when compared with various more recent features, such as spectral subband centroids (SSCs) (Kinnunen et al., 2007; Thian et al., 2004) they performed better. The main reason for that lies in the fact that MFCC unlike regular cepstrum uses frequency bands that are equally spaced on the mel scale, which approximates the human auditory system’s response more closely than the linearly-spaced frequency bands used in the normal cepstrum.

To extract MFCC from an audio signal, the signal is processed as shown in Figure 5. At the end of this pipeline, 13 coefficients were extracted. The 0th coefficient is discarded since it is a very narrow band and contains powers that exist near 0 Hertz which is not significant for the task. The delta and delta-delta were added to the remaining 12 coefficients to ensure that the model is getting sufficient information about the fluctuation of the signal.

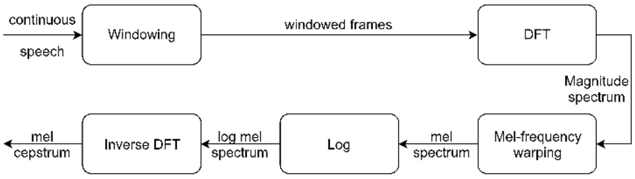


Figure 5. MFCC derivation

$$\Delta_k = f_k - f_{k-1} \tag{1}$$

$$\Delta\Delta_k = \Delta_k - \Delta_{k-1} \tag{2}$$

2.1.2 Linear Prediction Cepstral Coefficients

Linear prediction (LP) is another alternative for short spectrum analysis. It has a good intuition about the interpretation of both correlated adjacent samples (in the time domain) and resonant corresponding poles (in the frequency domain). The assumption here is that the entire speech process can be represented by a digital filter shown in Figure 6 which makes LP one of the most powerful type 2 features for speech analysis (Buza et al, 2006).

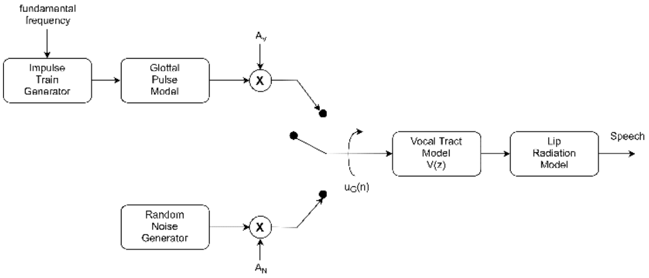


Figure 6. Speech Production Model. A_v is the voiced sound gain, A_n is the unvoiced sound gain and $u_G(n)$ is the voiced/unvoiced switching function

To extract LPC from an audio signal, the signal is first framed using short time Hamming windows. For each frame, 14 coefficients are extracted. This ensures that all possible speech segments (voiced and unvoiced) are covered and for both genders as well. Coefficients are extracted by computing the vector a_k that links the current speech sample with the previous samples with the same window. This can be expressed by the following equation:

$$s(n) = \sum_{k=1}^{p} \alpha_k . s(n-k) \qquad (3)$$

Where s is the speech signal, n is the sample point, α is the formant and p is the number of the required formants (14 in our case). LP coefficients are not often used for speaker recognition. Instead, the linear predictive cepstral coefficients (LPCCs) are incorporated since they are more robust and less correlated. LPCCs are estimated by applying DFT to the obtained LP coefficients.

2.1.3 Perceptual Linear Prediction.

PLP is another feature that describes the psychophysics of the human speech production system. As shown in Figure 7. The procedure for extracting PLP is similar to that of MFCC with the difference lie in the incorporation of Bark-spaced filterbanks instead of Mel-spaced filterbanks, making this feature perfect as a type 2 feature.

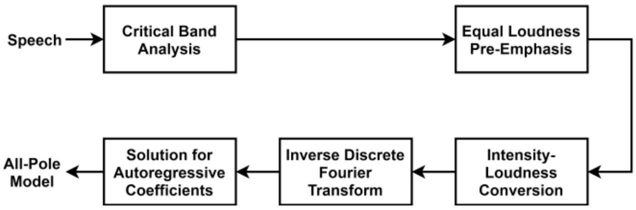

Speech
Critical Band Analysis
Equal Loudness Pre-Emphasis
Intensity-Loudness Conversion
Inverse Discrete Fourier Transform
Solution for Autoregressive Coefficients
All-Pole Model

Figure 7. PLP derivation

2.2 Data Collection

To carry out this work, sixteen participants were recruited from the Institute of Electrical and Electronic Engineering (IGEE) to record a corpus of 51 to 54 English sentences each. All participants were Ph.D. students with ages ranging from 24 to 27 (8 females, 8 males; mean age: 25.37 ± 0.88). They were all non-native English speakers from Algeria, with a relatively high level of proficiency in English (mean TOEFL score: 89.75 ± 6.75 with a mean of 26.06 ± 2.93 in the speaking section).

All recordings were carried out in the same room located within the institute, (6.0m(L) x 3.5m (W) x 4m(H) as shown in Figure 8). Participants were sitting on a chair, facing a wall at a distance of 0.75m with a monitor in front of them displaying the sentence to be read. The recording device (Honeywell CN51 Personal Digital Assistant (PDA)) was placed between the monitor and the participant.

The recorded sentences differ from one student to the other to ensure that the task remains text-independent and the collaboration of the speaker to a minimum.

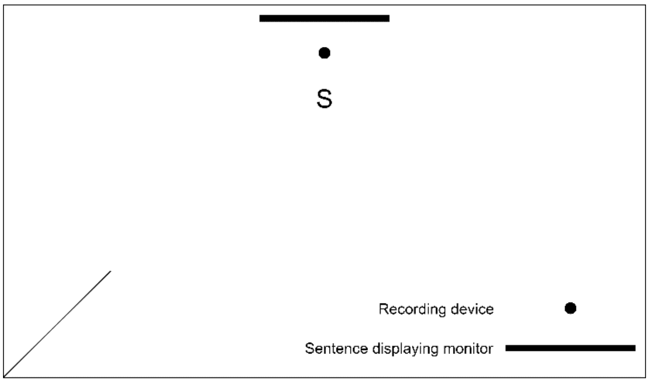


Figure 8. Recording room layout. S is the speaker/participant

To demonstrate that the approach is effective for a wider range of scenarios, the LibriSpeech ASR corpus (Vassil et al, 2015) was implied. The corpus however was only used to confirm the recognition accuracy and was not used for all procedures due to hardware limitations.

2.3 Procedures

To assess the performance of the feature fusion approach, a specified pipeline was put in place.

2.3.1 Tuning the Parameters for Feature Extraction

Contrary to what is known in the literature (Kinnunen and Li, 2010) the selection of the frame width and increment is important. These parameters need to be adjusted so that the changes are not too drastic but also capture the right information. In fact, (Paliwal et al, 2010. Eringis & Tamulevicius, 2014) demonstrated this in their works (Figure 9).

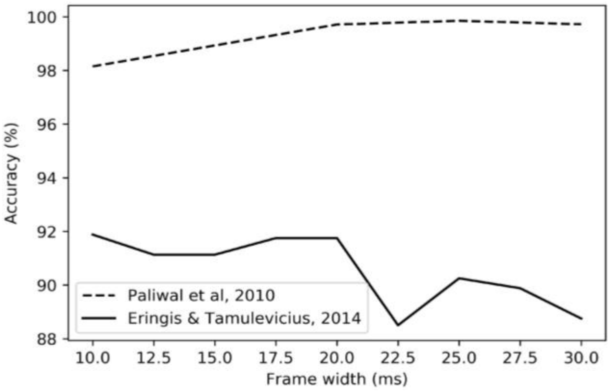


Figure 9. The effect of adjusting the frame width during MFCC extraction on the overall accuracy

There is some discrepancy in the literature regarding the frame width values which optimize accuracy (see Figure 9). To confirm which parameters to choose, the structure shown in Figure 10 is trained with each of the suggested features, extracted using different frame widths. Thirty random segments for each speaker in the data set are used for training while the remaining 21 segments are used for testing. This training was carried out in a cross-validation setup to avoid any overfitting due to the complex nature of the chosen model. Once the overall accuracy was obtained, the frame width was increased by 5 ms. This procedure was repeated until a frame width of 30.0 ms was reached.

The second parameter to consider for tuning the feature extraction was the frame increment. For this, 3 values were considered: Overlapping (50%), Slightly-Overlapping (75%), and Non-Overlapping (100%). For each of these values, the same network shown in Figure 10 was trained with 30 cross-validated speech segments.

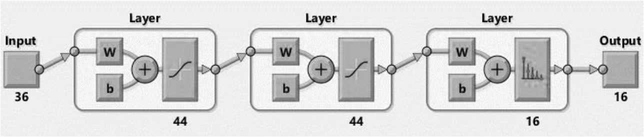


Figure 10. The single feature ANN structure

2.3.2 Fusing, Training, and Testing

To study the effect of fusing features on the overall recognition rate, the structure shown in Figure 2 is trained using two features extracted at the optimal frame width and increment. For each speaker, 30 random segments are used for training while the remaining 21 segments are used for testing. The selected segments were cross-validated and the resulting performance was compared to that of training the structure shown in Figure 10 using a single feature extracted at the same optimal parameters.

The same procedure was repeated for the LibriSpeech corpus. To ensure that the approach works on a larger dataset (921) and on native English speakers. Although the corpus contains more than 51 segments for each speaker, only 51 segments were used between training and testing similar to the recorded corpus.

2.3.3 Tokenizing using Advanced Neural Network

To validate the effectiveness of our approach (i.e. feature fusion) both structures shown in Figures 3, 4 were used as tokenizer for the newly developed feature. Similarly to 2.3.2, the structure in Figure 2 was trained using a pair of type 1 & 2 features. However, instead of using all of the provided speakers, 12.5% (2 speakers for the recorded corpus and 115 speakers for LibriSpeech) of that is used to obtain the pre-trained fusion network (Figure 2) that would define our mapping from regular features to speaker-specific ones.

With the new mapping defined Figures 3, 4 are trained and tested in a similar fashion as in section 2.3.2 (i.e. 30/21 segments with 10 fold cross-validation) using the remaining 87.5% speakers from each data set. Finally, the obtained accuracy from these newly defined features as compared to that of the same structures trained using each of the suggested features (MFCC, LPCC, and PLP).

Given the shape of the input for CNN, all implied segments for this step were kept at a fixed duration of 6 seconds (i.e. either cropped or zero-padded).

3. RESULTS & DISCUSSION

When adjusting for the frame width during feature extraction, the best recognition rate was obtained for a frame width of 10 ms, which coincides with the results obtained by Eringis and Tamulavicius, 2014. These results are depicted in Figure 11. This pattern of improvement was found across all three features that were extracted.

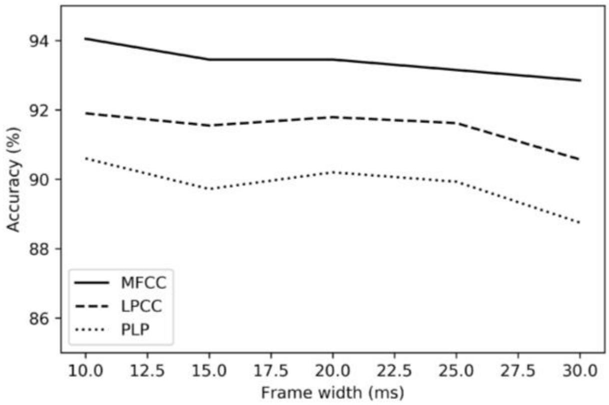


Figure 11. The results of frame width tuning tasks

Examining the 3 different frame increments for the optimal frame width of 10 ms, we found that a frame increment of 75% provided the highest recognition rate on average compared to the 50% and 100% increments (see Figure 12). This implies that by only adjusting the extraction parameters, 3 additional speech segments were correctly identified.

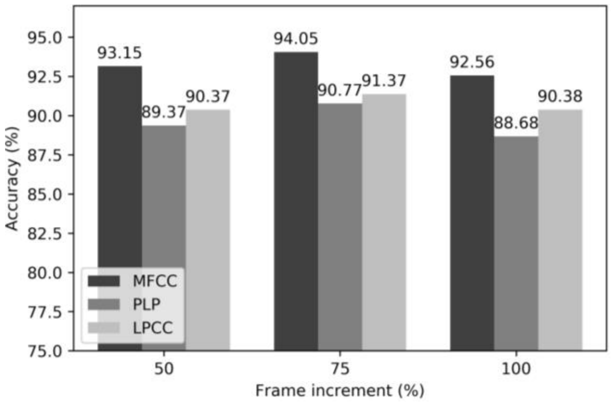


Figure 12. The results of the frame increment tuning task

The fusion of type 1 and type 2 features improved the accuracy of the recognition tasks. As seen in Figure 13, any combination of type 1 and type 2 features would result in a better recognition rate compared to using a single feature. The best result was obtained when combining MFCC and LPCC (99.4% accuracy), this is due to the fact that PLP is more optimized for speech recognition tasks and not speaker recognition tasks (Hermansky, 1990). This in fact explains the low recognition rate in all of the proceedings results.

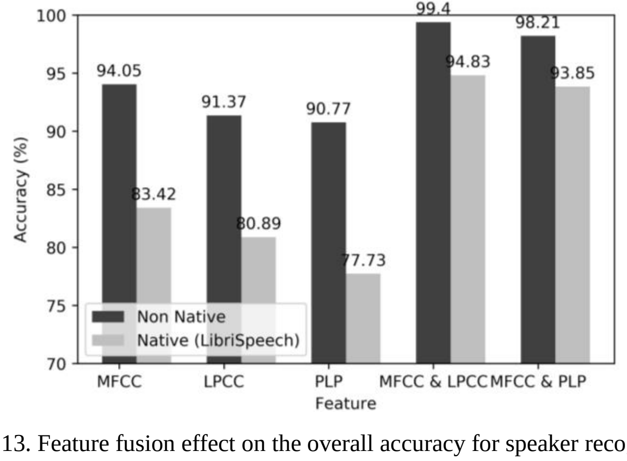


Figure 13. Feature fusion effect on the overall accuracy for speaker recognition

Figure 13 also demonstrates that the approach works well for native speakers and for a larger set of speakers. In fact, the improvement obtained when testing for the larger corpus was higher. This was due to the fact that this approach created higher dimensionality allowing for better discrimination. Of course, the approach resulted in a slight increase in training time. Figure 14 shows the required training time for each of the trials described in section 2.3.2. These results were obtained using the 4710MG i7 CPU with a RAM of 8GB and only for the recorded corpus.

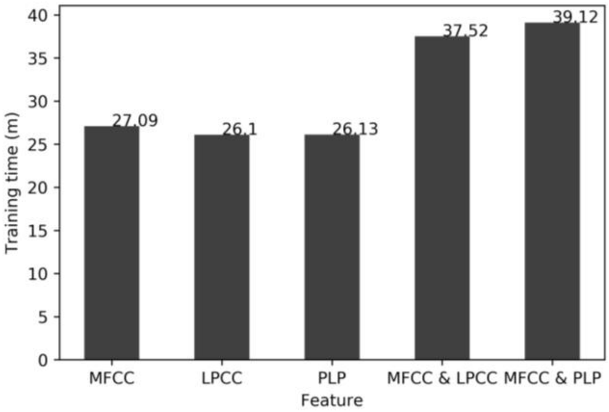


Figure 14. Feature fusion effect on the required training time for speaker recognition

The difference in training time between the best performing single feature structure and the best performing fused features structure is 10 minutes and 25 seconds. This means an increase of 38.50% in training time for an improvement of 5.35% in speaker-recognition accuracy if we consider only the effect of feature fusion without the parameter tuning and 6.55% when considering both tweaks. The training time can be significantly reduced if better hardware is used such as a more performing GPU or/and by reducing the duration of the recordings that are used for training the model as feature fusing allows to recognize the speaker much earlier as shown in Table 1.

Table 1. Speaker recognition accuracy over time

Input	**Time (s)**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MFCC	11.01	13.09	19.94	22.91	27.97	50	86.9	91.07	92.85	94.05
LPCC	10	12.5	19.34	22.91	27.38	39.88	83.33	88.89	89.28	91.37
PLP	10.5	12.79	19.94	21.72	27.08	39.58	81.84	86.60	88.39	90.77
MFCC&LPCC	12.20	18.15	32.14	63.09	84.82	**90.17**	**94.34**	97.91	98.57	99.4
MFCC&PLP	12.20	17.85	31.54	60.11	77.08	85.11	90.17	95.23	96.72	98.21

By using 6 seconds of recording the model was able to reach 90.17% when fusing MFCC with LPC. What is even more important about these results is the fact that the fusion approach was able to outperform the single feature approach by utilizing 7 seconds out of the provided 10 seconds of recording, this reduced training time in the fusion approach from 37 minutes and 32 seconds to 24 min and 41 seconds. This time is less than the time required to train the model for any of the features independently. In fact, this observation is the reason why the LSTM and CNN (Figures 3, 4) were trained using only 6 seconds of speech. The use of transfer learning to define a speaker-specific mapping for two streams of features worked and yielded in a huge boost the recognition rate as shown in Figure 15.

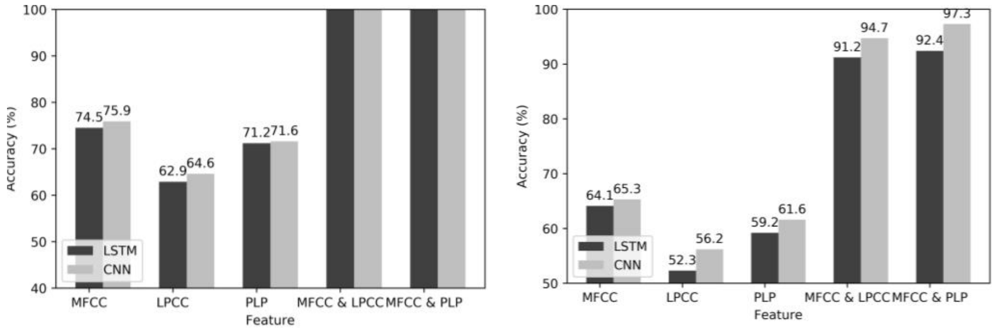


Figure 15. The results of using the transferred feature on the overall accuracy, left is for the recorded corpus, and right is for LibriSpeech corpus

Although the accuracy of the individual feature increased due to the use of advanced neural networks, it's still not sufficient enough to be considered for critical application. However, the features defined through transfer learning had a huge boost in performance leading even to saturation (100% accuracy) for the smaller recorded corpus. In addition, to the huge boost in performance by pre-training the mapping function beforehand, the training time for fused features or for a regular acoustic feature, in this case, is identical.

What's worth noting is the fact that unlike the previous results, PLP performed better on its own and also when paired with MFCC to recreate the new feature. We believe that this is mainly due to the fact PLP although are less speaker-specific on a frame to frame basis, has the potential of providing clearer tokens for recognizing co-occurring patterns of phonetic information.

4. CONCLUSION

The speech signal does not only convey a message, it conveys information about the speaker themselves, their gender, origins, health, and age. The aim of this work was to improve the task of recognizing a person based on speech segments.

In this work, we set to redefine what a speaker-specific feature should really be, and how to extract them. The proposed approach relies on ANN and their ability to extract intermediary features that are transferable from task to task. Based on this, a structure was trained using two sets of features providing complementary information (type 1 & 2 features) to define a mapping function from the space of these parallel streams to a space of speaker-specific features. The approach proved to be indeed effective and yielded a ~33% increase in the recognition rate. We believe that such an approach may revolutionize what a speaker-specific feature should be? and instead of using acoustic features designed for speech recognition applications, we can either remap them to a more suited feature space or use raw data such as FFT or wavelet to extract an entirely new feature.

The only drawback of such an approach is the additional training time required to train the mapping neural network. Nevertheless, this can be mitigated if sufficient numbers of speakers are implied.

REFERENCES

Al-Shayea Q.K., and Al-Ani M.S., 2016. Speaker Identification: A Novel Fusion Samples Approach. *In International Journal of Computer Science and Information Security*, Vol. 14, No. 7, pp 423-427.

Andrews W., Kohler M., Campbell J., Godfrey J., Hernandez-Cordero J., 2002. Gender-dependent phonetic refraction for speaker recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Orlando, Florida, USA, pp. 149–152.

Awais M., Mansour A., and Ghulam M., 2014. Automatic Speaker Recognition Using Multi-Directional Local Features (MDLF). *In Arabian Journal for Science and Engineering*, Vol. 39, No. 5, pp 3379-3811.

Brooks M., 1997. Voicebox: A Speech Processing Toolbox for MATLAB.2006. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Buza O., Toderan G., Nica A., and Caruntu A., 2006. Voice Signal Processing for Speech Synthesis, *International Conference on Automation, Quality and Testing, Robotics*. Cluj-Napora, Romania, pp. 360-364.

Campbell W., Campbell J.P., Reynolds D.A., Jones D., Leek T., 2004. Phonetic speaker recognition with support vector machines. *In Advances in Neural Information Processing Systems*, MIT Press, Vancouver, British Columbia, Canada.

Campbell W.M., Campbell J.P., Reynolds D.A., Singer E., and Torres-Carrasquillo P.A., 2006. Support vector machines for speaker and language recognition. *In Computer Speech and Language*, Vol. 20, No. 2,pp 210-229.

Chakroborty S. and Saha G., 2009. Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter. *In International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering,* Vol. 3, No. 11, pp 1974-1982.

Chakroum R., Zouari L.B., Frikha M., and Ben Hamida A., 2016. Improving Text-independent Speaker Recognition with GMM. *International Conference on Advanced Technologies for Signal and Image Processing,* Monastir, Tunisia, pp. 693-696.

Charbuillet C., Gas B., Chetouani M., Zarader J., 2006. Filter bank design for speaker diarization based on genetic algorithms. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Toulouse, France, pp. 673–676.

Davis S., Mermelstein P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *In IEEE Transactions on Acoustics, Speech, Signal Processing,* Vol. 28, No. 4,pp 357– 366.

Donald R., 1985. Conversational voice command control system for home appliance. *In U.S Patent*, n. US4520576A.

Eringis D., and Tamulevicius G., 2014. Improving Speech Recognition Rate through Analysis Parameters. *In Electrical Control and Communication Engineering,* Vol. 5, No. 1, pp 61-66.

Gaafar T.S., Abo Baker H.M., and Abdalla M.I., 2014. An improved method for speech/speaker recognition. *International Conference on Informatics, Electronics & Vision*, Dhaka, Bangladesh, pp. 1-5.

Hermansky H., 1990. Perceptual linear prediction (PLP) analysis for speech. *In The Journal of the Acoustical Society of America,* Vol. 87, No. 4,pp 1738–1752.

Huang X., Acero A., Hon H.-W., 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice-Hall, New Jersey, USA.

Kinnunen T., 2002. Designing a speaker-discriminative adaptive filter bank for speaker recognition. *Proceedings of the International Conference on Spoken Language Processing (ICSLP),* Denver, Colorado, USA, pp. 2325–2328.

Kinnunen T., Zhang B., Zhu J., Wang, Y., 2007. Speaker verification with adaptive spectral subband centroids. *Proceedings of the International Conference on Biometrics (ICB),* Seoul, Korea, pp. 58–66.

Kinnunen T., Li H., 2010. An overview of text-independent speaker recognition: From features to supervectors. *In Speech Communication*, Vol. 52, No. 1,pp 12-40.

Makhoul J., 1975. Linear prediction: a tutorial review. *In Proceedings of the IEEE,* Vol. 64, No. 4, pp 561–580.

Mammone R., Zhang X., Ramachandran R., 1996. Robust speaker recognition: a feature based approach. *In IEEE Signal Processing Magazine*, Vol. 13, No. 5,pp 58–71.

Meena K., Subramaniam K., and Gomathy M., 2011. Gender Classification in Speech recognition using Fuzzy Logic and Neural Network. *In The International Arab Journal of Information Technology,* Vol. 10, No. 5,pp 477-485.

Miyajima C., Watanabe H., Tokuda K., Kitamura T., Katagiri S., 2001. A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction. *In Speech Communication*, Vol. 35, No. 3-4,pp 203–218.

Murty, K., Yegnanarayana, B., 2006. Combining evidence from residual phase and MFCC features for speaker recognition. *In IEEE Signal Processing Letter,* Vol. 13, No. 1,pp 52–55.

Nolan F., 1983. *The Phonetic Bases of Speaker Recognition*. Cambridge University Press, Cambridge, United Kingdom.

Orman D., Arslan L., 2001. Frequency analysis of speaker identification. *Proceedings of the Speaker Odyssey: The Speaker Recognition Workshop (Odyssey),* Crete, Greece, pp. 219–222.

Paliwal K.K., Lyons J.G., and Wojcicke K.K., 2011. Preference for 20-40 ms window duration in speech analysis. *International Conference on Signal Processing and Communication Systems,* Gold Coast, Australia, pp. 1-4.

Singh S., Assaf M.H., Das S.R., Biswas S.N., Petriu E.M., and Groza V., 2016. Short Duration Voice Data Speaker Recognition System Using Novel Fuzzy Vector Quantization Algorithm. *International Instrumentation and Measurement Technology Conference Proceedings,* Taipei, Taiwan, pp. 1-6.

Sinno J.P., Qiang Y., 2010. A Survey on Transfer Learning. *In IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 10,pp 1345–1359.

Slaney M., 1998. Auditory Toolbox. [Online]. Available: https://engineering.purdue.edu/~malcolm/interval/1998-010/

Srinivas V., Santhi C.R. and Madhu T., 2014. Neural Network based Classification for Speaker Identification. *In International Journal of Signal Processing, Image Processing and Pattern Recognition,* Vol. 7, No. 1, pp 109-120.

Statista Research Department, 2020. Number of digital voice assistants in use worldwide 2019-2023. [Online]. Available: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

Syntellect Inc., 2003. The Importance of Creating an Effective Voice User Interface. [Online]. Available: https://www.contactcenterworld.com/view/contact-center-article/the-importance-of-creating-an-effective-voice-user-interface.aspx

Thian N., Sanderson C., Bengio S., 2004. Spectral subband centroids as complementary features for speaker authentication. *In Proceedings of the International Conference on Biometric Authentication (ICBA),* Hong Kong, China, pp. 631–639.

Vacic V., 2015. Summary of the training functions in Matlab’s NN toolbox. *In MSc, University of California,* Riverside, California, USA.

Vassil P., Guoguo C., Daniel P., and Sanjeev Khudanpur., 2015. LibriSpeech: an ASR corpus based on public domain audio books. *International Conference on Acoustics, Speech, and Signal Processing,* South Brisbane, Queensland, Australia, pp. 5206-5210.

Wolf J. J., 1972. Efficient Acoustic Parameters for Speaker Recognition. *In Journal of The American Statistical Association*, Vol. 51, No. 6,pp 2044–2056.

Ye J., Yu Z., Ron J. W., Quan W., Jonathan S., Fei R., Zhifeng C., Patrick N., Ruoming P., Ignacio L. M., and Yonghui W., 2018. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *In Advances in Neural Information Processing Systems*, Vol. 31, No. 1,pp 4485-4495.

Zheng N., Lee T., Ching P., 2007. Integration of complementary acoustic features for speaker recognition. *In IEEE Signal Processing Letters,* Vol. 14, No. 3, pp 181–184.