GSoC 2015 Proposal : Identify male and female speakers

Brij Mohan Lal Srivastava

International Institute of Information Technology, Hyderabad

brijmohanlal.s@research.iiit.ac.in

http://researchweb.iiit.ac.in/~brijmohanlal.s/

Abstract

Splitting an audio input based on gender of speaker is generally a hard problem. But there have been many significant works that have been proved to be very effective. This goal of this project is to temporally segment an input file, based on whether it is spoken by a male or female.

In order to do that, we will evaluate the performance of state-of-the-art tools and apply some of the speech features, that we feel might increase the efficiency of the task.

1. Introduction

Speech is a very interesting since it contains a lot of information about the personality of an individual. A soon as we hear someone, we assume the gender, age group, mood and such information related to the speaker apart from actual information that he/she want to convey. Detecting gender is very easy for humans, because of the pitch variations between male and female. But how do we learn that threshold and apply it in our algorithms?

Rouvier et al. [1] presents LIUM open-source speaker diarization toolbox. It uses advanced Hierarchical Agglomerative Clustering techniques and performs very well on Broadcast News. Simpson and Adrian [2] study the phonetic differences between male and female speech. They suggest some production-based features, like interaction of pitch and articulation, VOT, duration, etc. which can be used to enhance the performance, or cross-validate the output of state-of-the-art systems like LIUM [1] and Shout [1].

In a huge corpus as available in case of Redhen Labs, we may encounter several such instances when the same speaker recurs again in a different recording. It will be highly intuitive if we can identify such instances and annotate the segment with the speaker already encountered in previous audio recordings. Fortunately, LIUM [1] provides such a functionality by means of CLR (Cross-Likelihood Ratio) and ILP (Integer Linear Programming) clustering techniques [2]. I shall be utilizing the BIC (Bayesian Information Criterion) clustering provided as a module in LIUM, to cluster the same speakers across the corpus. Implementation section contains detailed information about this approach.

Redhen lab works on data collected which mainly consists of television serials and news. This is a huge dataset and need to be annotated with various meta-information. Speaker- or Gender-based segmentation is highly improtant step towards giving completing this meta-information, which I would like to achieve by using available tools and literature.

[1] http://shout-toolkit.sourceforge.net/use_case_diarization.html

[2] http://www-lium.univ-lemans.fr/diarization/doku.php/cross-show_diarization

2. Project Goals

The main aim of this project is to achieve best possible gender-based segmentation of speech. This should happen in real-time scenarios where there can be:

1. Background noise
2. Cross-talk

This is the rough walk-through of my approach:

1. Explore state-of-the-art tools, like LIUM and Shout
2. Prepare an accurate ground-truth
3. Measure the accuracy of tools against ground-truth, in $noisy$ and $crosstalk$ conditions
4. Cluster the speakers across the corpus based on BIC Clustering technique
5. Find the areas where there is mis-match and their causes
6. Post-process the output by re-inforcing it with other relevant information like, pitch, acoustic characteristic at point of mis-match
7. Create a framework to incorporate end-to-end segmentation with all the post-processing in pipeline
8. If possible, write a tool from scratch specifically for male-female segmentation.

3. Implementation

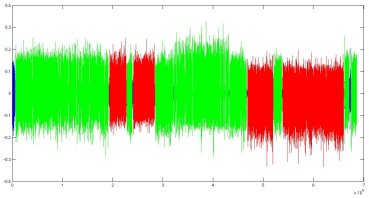
Here I will present some of my preliminary findings/experimentations and approach.

3.1. Observations

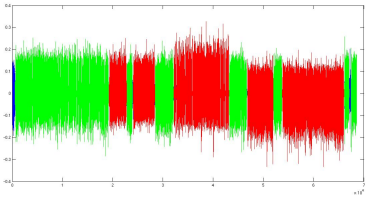
A sample audio file was manually annotated for male/female segmentation. The same file was given to LIUM for gender classification. Figure 1 shows the result. LIUM performed very well for the segments which were prominently distinguishable based on F0, but we can see a mis-classified segment when the pitch of male speaker is unusually high and overlaps with female pitch region.

As seen in Figure 2, the distribution of pitch mass is towards higher regions when there's a female speaker and low regions otherwise. This information can help us refining the output of LIUM. We can pick the segments provided by LIUM and cross-verify them with the learnt value of pitch. The learnt value can be statistically obtained or can be learnt by using deep learning.

A regular moving average of pitch value (Figure 3) also gives us fairly good information, which help us get almost 65% accuracy by just using that information. This can be a good additional feature in learning the distinction.



(a) Manually annotated ground-truth



(b) Segments annotated by LIUM

Figure 1: Comparison of LIUM output with ground-truth

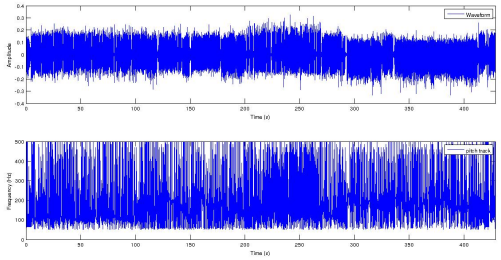


Figure 2: *Pitch track of sample audio*

3.2. Apporach

We can extend the current approach of applying speaker diarization to single audio file. This technique will be executed in two passes. In the first pass, we will process single audio files sepearately. Then Hierarchical Agglomerative Clustering with BIC as distance measure can be applied to the output of first pass in order to merge the diarization results based on same speaker.

Due to the advent of deep learning, its fairly obvious to incorporate it for the current task. The challenging part with classification networks is the dataset. They require hige training and testing data in order to draw generic boundaries. In order to obtain such data, we can bootstrap by annotating the dataset using LIUM + post-processing. We can then utilize this dataset to train the network. This approach might result in better performing classifiers. Recently, convolutional neural networks have shown that they can extract highly discriminatory features from speech [3] [4] [5] [6] [7]. We can harness this property of CNN to classify audio segments into male/female and build an in-house tool. Figure 4 shows a block diagram of the flow of this approach roughly.

4. Timeline

The above approach can be sub-divided into small do-able chunks:

During community bonding

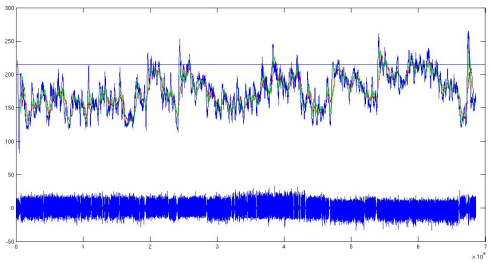


Figure 3: *Moving average of audio pitch*

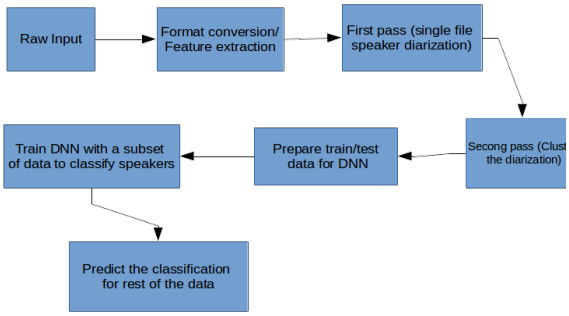


Figure 4: *Approach*

1. **Apr end - May 15** : Exploring strong points of tools like $Praat$, $LIUM$, $Shout$, $Idiap$, etc. along with deep learning literature for speaker diarization.
   (a) Prepare a strong ground-truth manually or with minimum error which can be used to evaluate system's accuracy
   (b) Each tool will be tested against a subset of data for which we hold a strong ground-truth
2. **May 16 - May 26** : Observe the waveform and go through the literature to find any relevant features which can help disambiguate if there are close-values of confidence measure over segments. These are the spurious segments and need to be resolved in order to get better accuracy.

Post community bonding

1. **May 27 - June 15** : Implement a post-processing module
   (a) Convert each file to appropriate audio format (mostly 16k Hz 16-bit wav)
   (b) Create programs in Python/bash to run first pass of speaker diarization on each file and collecting speaker and gender Gaussian Mixture Models alongwith diarization output.
   (c) Run the second pass to cluster/consolidate the diarization results across the corpus (Python/bash)
2. **June 15 - June 20** : Create training/testing dataset for next phase by using the system built till now
   (a) Analyze the data to select diverse data for training and testing
   (b) Create scripts for segregating the dataset

3. **June 21 - June 30** : Experiment with various architectures of various Deep Neural Networks, like Recursive NN, Recurrent NN, Convolutional NN, etc. and go through literature to find the best network suitable for this task.
    (a) Setup environment for running DNN, like Torch7 or Theano
    (b) Prepare code samples in Lua or Python for running same experiment of predicting speaker classes in different DNNs (RNN, CNN, etc)
    (c) Report the speed/accuracy trade-offs in each architecture for the current task
    (d) Decide on the best network which will be further used
4. **July 1 - July 31** : Code the neural network needed for this task and test with various samples
5. **Aug 1 - Aug 20** : In-house framework development
    (a) Create an end-to-end framework which takes audio file as input, extracts relevant features, preprocess (if needed), feed the features to NN, obtains the output and post-process in a standard format.

5. About Me

I am a Research student in Speech and Vision Lab, at International Institute of Information Technology- Hyderabad, India.

I am a working under Dr. Manish Shrivastava. My area of focus is Spoken Dialog Systems. I am working on creating a domain- independent virtual assistant for rural population as a part of my thesis. I have worked on speaker identification using GMMs as part of Speech Technology course conducted by Dr. Kishore Prahallad. I have also attended Speech Signal Processing course taught by Dr. Yegnanarayana which has helped me to get insight on basic speech concepts.

Kindly visit my webpage for more information (http://researchweb.iiit.ac.in/~brijmohanlal.s/)

6. References

[1] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," Idiap, Tech. Rep., 2013.

[2] A. P. Simpson, "Phonetic differences between male and female speech," *Language and Linguistics Compass*, vol. 3, no. 2, pp. 621–640, 2009.

[3] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.

[4] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[5] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.

[6] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Proceedings of the 15th Conference of the International Speech Communication Association, Interspeech*, 2014.

[7] D. Palaz, M. M. Doss, and R. Collobert, "Learning linearly separable features for speech recognition using convolutional neural networks," *arXiv preprint arXiv:1412.7110*, 2014.