Interspeech 2018
2-6 September 2018, Hyderabad

Speaker Recognition with Nonlinear Distortion: Clipping Analysis and Impact

Wei Xia, John H. L. Hansen

Center for Robust Speech Systems
University of Texas at Dallas, Richardson, TX 75080
wei.xia@utdallas.edu john.hansen@utdallas.edu

Abstract

Speech, speaker, and language systems have traditionally relied on carefully collected speech material for training acoustic models. There is an overwhelming abundance of publicly accessible audio material available for training. A major challenge, however, is that such found data is not professionally recorded, and therefore may contain a wide diversity of background noise, nonlinear distortions, or other unknown environmental based contamination or mismatch. There is a critical need for automatic analysis to screen such unknown data sets before acoustic model development, or to perform input audio purity screening prior to classification. In this study, we propose a waveform based clipping detection algorithm for naturalistic audio streams and analyze the impact of clipping at different severities on speech quality measures and automatic speaker recognition systems. We use the TIMIT and NIST SRE-08 corpora as case studies. The results show, as expected, that clipping introduces a nonlinear distortion into clean speech data, which reduces both speech quality and speaker recognition performance. We also investigate what degree of clipping can be present to sustain effective speech system performance. The proposed detection system, which will be released, could contribute to massive new audio collections for speech and language technology development.

Index Terms: Clipping, Speaker recognition, Non-linear distortion

1. Introduction

The formulation of advanced speech and language technology in the past has historically relied on carefully organized data collection, typically in controlled laboratory conditions. However, today a vast amount of audio data appears daily on the web, data servers for call centers, and personal devices such as smart phones. Due to the cost in collecting organized and focused speech/language corpora, researchers have turned to found naturalistic audio material in order to reduce both cost and collection time, as well as increase the diversity of speakers, languages, and topic content. While there is a plethora of audio material available, non-uniformity and recording impurities could raise great concerns regarding the viability of resulting algorithms.

Audio peak clipping occurs when the volume of the recorded audio signal exceeds the input range of the microphone's pre-amplifier, or the audio data is recorded without an appropriate input automatic gain control (AGC). Portions of the signal above the maximum voltage would be clamped to the maximum value of the signal when it passes through the analog to digital (A/D) converter. The loss of high amplitude samples introduces a non-linear distortion in the form of odd harmonics in high frequencies, resulting in audible artifacts in the recorded audio.

In order to detect such distortions, Aleinik et al. [1, 2] presented a histogram method to estimate the level of signal clipping. Ding et al. [3] investigated the effects of temporal clipping on perceived speech quality. They proposed a non-intrusive algorithm based on clipping statistics to predict speech quality. Temporal clipping can also occur as a result of voice activity detection (VAD) or echo cancellation where comfort noise is used in place of clipped speech segments. Eaton and Naylor [4, 5] proposed a Least Squares Residuals Iterated Logarithm Amplitude Histogram (LILAH) based approach for detecting clipping in speech that shows robustness to the speaker, clipping level, and codec applied, and provides an estimate of the original signal level. It could achieve great performance without prior knowledge of the encoding used in the original signal. Bie et al. [6, 7] proposed a distribution based clipping detection algorithm and a signal reconstruction approach based on deep neural networks. Tachioka et al. [8] analyzed the relationship between clipping level and automatic speech recognition performance and showed an explicit relationship between SNR and clipping level.

To quantify the clipping impact on speech quality, Hines et al. [9, 10] presented a non-reference measure that uses a modular design to help pinpoint the reason for degradations. Maymon et al. [11] proposed two iterative approaches based on the band-limited assumption and auto-regressive model to recover clipped speech signal. Harvilla and Stern [12, 13] introduced a de-clipping algorithm based on constrained least-squares minimization, where the constrained blind amplitude reconstruction algorithm interpolates missing data points such that the resulting function is smooth while ensuring the inferred data falls into a legitimate range.

Given the ever increasing availability of speech and audio data in the field, the speech community needs improved speech tools to better characterize and understand issues in found data. In this study, we restrict the potential acoustic issue to waveform peak clipping, which occurs when audio data is recorded at an improperly adjusted gain setting and without an input AGC, or when the AGC cannot respond quickly enough to suppress impulsive audio events. Clipping detection can be even more challenging when clipped data is transformed to different formats (e.g., wav, mp3, etc), since peak clipped values can be re-assigned to new clipped max values.

In the following sections, we first give an example of speech signal clipping effect in the temporal, spectral, and cepstral domains respectively. Next, we present an algorithm to detect and tag audio waveform peak clipping in Section 3. In Section 4, we show the overall impact of clipping on automatic speech systems by conducting speech quality measurement analysis and investigating the impact of clipping on automatic speaker recognition. We also discuss the specifics of the proposed algorithm used for this analysis, which will be distributed for general usage to the community. Finally we conclude in Section 5 with

10.21437/Interspeech.2018-2430

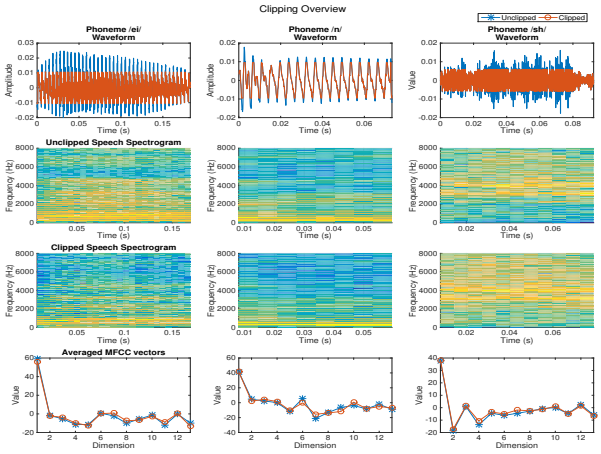

Figure 1: *Comparison of original clean and clipped speech that includes waveform, spectrogram, and averaged MFCC feature vectors for three phonemes (/ei/, /n/, /sh/)*

future work.

2. Clipping Overview

Speech peak clipping occurs when the volume of the speech signal being recorded exceeds the input voltage range of the microphone's pre-amplifier given the current gain for analog-to-digital (A/D) conversion. When this occurs, the pre-amplifier voltage becomes saturated, and unable to provide an accurate discrete representation for reliable A/D conversion. This causes the peak of the signal to not be reproduced by the pre-amplifier, and all portions of the signal above the maximum voltage of the pre-amplifier will be clamped to the maximum value as it passes through the A/D converter. The manifestation of this loss of data, and the introduction of this plateau shape comes in the form of non-linear distortion, especially odd harmonic distortion in higher frequencies, resulting in audible artifacts in the recorded audio.

We illustrate an example of the effects of clipping for speech signals in Fig. 1. The first row presents time domain waveforms of the /ei/, /n/, and /sh/ phonemes from left to right respectively, with blue representing the original unclipped signal, and red the corresponding clipped version. We create these visuals by isolating the three phonemes and increasing the gain of each until 10% of the audio time domain samples are clipped.

We show spectrograms of the original unclipped waveforms in the second row, versus clipped waveform spectrograms in the third row. Comparing the corresponding spectrogram plots of each phoneme, we observe that the largest impact of clipping is distortion harmonics appearing in the higher frequencies of the spectrogram. This manifests as more energy within the upper regions of the spectrogram. We note that the lower frequencies, and the overall shape of the speech energy information still remains, and the formant structure appears to be present and intact. This would imply that the audio is still intelligible, and that, for this level of clipping, the content is not degraded to the point that it is very difficult to understand. Our subsequent listener test also supports this observation, with the speakers information content is very clear, but with noticeable high frequency artifacts present such as pops and hisses. The fourth row of Fig. 1 displays the averaged mel-frequency cepstral coefficients (MFCC) plots for the same three pairs of waveforms. These plots represent a small slice of the data that would be processed and used by classifiers in automatic speech-processing systems to represent speakers or phonemes. Comparisons across these plots show how different clipped versus unclipped files would be interpreted by automatic speech systems. The overall shape of the clipped-audio MFCC vectors remains relatively similar to their unclipped counterparts. However, there is sufficient variation in individual coefficients which could impact the performance of speech systems. In Section 4, we will quantify the impact of clipping on speech quality assessment and speaker recognition systems.

3. Clipping Detection and Simulation

We propose a Clipping Detection and Tagging (ClipDaT) algorithm to detect signal clippings for naturalistic audio streams. Signal clipping may occur at the maximum level of the output digital representation. It may also occur when the speech signal is re-transmitted or converted to other audio file formats, where a renormalization of the overall gain is introduced after clipping.

The ClipDaT algorithm searches for strings of consecutive audio samples whose amplitudes reach a specific +/- maximum sample found in the file, or whose amplitudes come very near to this +/- maximum. We describe the clipping detection procedure in the following pseudo-code:

Algorithm 1 Sequential Clipping Detection

Let $x(n), n = 0, 1, ..., N-1$ be the initial discrete time signal
for $i = 0 \rightarrow N-1$ samples **do**
  Find the maximum absolute value of the signal x_{max}
  Find all M peaks within a very small range $x_{max} \pm \epsilon$ of the speech amplitude distribution, denote peaks as $\{M_0, M_1, ..., M_M - 1\}$
  Estimated clipping threshold θ is given by

$$\theta = \frac{1}{M-1} \sum_{i=0}^{M-1} M_i$$

  If x(n) $\geq \theta$ or x(n) $\leq$-θ
    counts $\leftarrow$ counts + 1
  Clipping rate $\gamma = \frac{counts}{N}$
end for

The first step is to perform an initial pass through the audio file to determine where the +/- maximum sample values are located. During a second pass, the algorithm searches for these +/- maximum values. Once one of these extreme samples is encountered, the immediately succeeding samples are analyzed in turn to decide if the extreme sample is the start of a string of clipped samples, or an isolated peak itself. If two consecutive extreme samples are found, we tag it as a clipping event. For the following samples to be deemed part of this clipping event, they must remain within a prescribed range of the extreme value. We set the clipping threshold by computing the average of speech signal peak values. We adopt this approach after analyzing a number of clipped speech files and discover that a high amplitude input that causes clipping may not result in a steady and flat waveform. We observe some slight variations in amplitude values once the pre-amplifier has been saturated.

To analyze the clipping effect on speech data in Section 4, we simulate the clipping distortion in two steps: 1) Detect the

clipping rate of an audio file using our proposed ClipDaT algorithm; 2) If the original clipping rate is lower than the expected clipping rate, we iteratively reduce the clipping magnitude with a constant small step size until it reaches the desired clipping rate. If it is higher, we discard that audio file.

4. Impact of Clipping

This section presents two studies on the impact of speech signal clipping. The first investigates how speech quality varies due to the introduction of clipping to the recorded speech data. The second analyzes how automatic speaker recognition system is affected by the clipped data.

4.1. Subjective Speech Quality Test

We first perform a human listening test as a screening test for the presence of clipping distortion. In some cases, distortion may not be perceptually noticeable, however it may still impact speech system performance. In order to quantitatively measure how clipping distortion affects human perception of audio signals, we conduct a subjective human listening test.

We recruit 15 participants from the University of Texas at Dallas. They range in age from 20 to 35 years old, and include a mixture of native American English (AE) speakers as well as non-native AE speakers. None possessed any history of hearing loss.

We present all listeners with twenty audio files with various levels of clipping, and ask them to rate each on a mean opinion score (MOS) scale. Listeners evaluate each audio using an MOS scale range from 1 to 5, where each value is defined as follows: (1) Bad: very annoying, (2) Poor: annoying, (3) Fair: slightly annoying, (4) Good: Clipping/distortion is perceptible, but not annoying, (5) Excellent: Clipping/distortion is imperceptible. We present the same 20 files in the same order to each listener. The distribution of clipping throughout the 20 files is as follows; there are 4 each of 5 different levels of clipping: clean/unclipped, 0.5%, 1.0%, 5.0%, 10.0% and 15.0%. We show results in Table 1, which reflect the average opinion scores across all participants.

Clipping Rate(%)	Quality Measure	
	MOS	PESQ
0	4.6	4.1
0.5	4.5	3.9
1	4.2	3.7
5	4.0	3.1
10	3.8	2.7
15	3.0	2.1

Table 1: *Comparison of speech quality measures and human perceptions of different levels of clipped audio.*

From the results in Table 1, it is clear that human listeners experience a degrading speech quality when more speech samples are clipped. We also compare MOS scores with an objective speech quality measure PESQ to analyze their correlation. Human listeners might be more generous with speech quality ratings than PESQ. Listeners are required to select whole-number score denominations (e.g., 1,2,3,4,5), unlike PESQ which reports scores with a much higher degree of granularity. Each data point in the PESQ column of Table 1 represents an average of 6300 samples, while each entry in the human listener MOS column consists of 60 averaged scores. The Pearson correlation between MOS an PESQ is 0.96, which is statistically significant at an alpha level of 0.05. The high correlation between human perceptions and objective quality measures shows that speech perceptual quality would decrease consistently if we introduce more contaminated clipped data. We also observe that listeners do perceive the quality gap between 1% clipped and 5% clipped audio less clearly than the difference between the 10% and 15% clipping levels.

4.2. Objective Speech Quality Assessment

Next, we use four speech quality measures to evaluate the clipped speech data. The first, Perceptual Evaluation of Speech Quality (PESQ) [14] is a full reference measure derived from the ITU-T P.862 recommendation regarding speech quality for telephony applications. The PESQ speech quality compares the perceptual difference between clean and degraded signals. The output from the PESQ algorithm is a mean opinion score (MOS) from -0.5 to 4.5, with 4.5 being the highest score, representing excellent quality.

The speech-to-noise-ratio (STNR) algorithm developed by NIST (NIST-STNR) estimates the distance between the speech and noise content of a bi-modal histogram distribution of an audio file. It is a non-reference quality measure and higher values represent a better quality measurement.

Waveform amplitude distribution analysis signal-to-noise ratio (WADA-STNR) is the third quality measure evaluated. Kim and Stern [15] proposed WADA-STNR as a non-reference quality measure. This approach tries to model the amplitude distribution of a speech waveform with the Gamma distribution function, while assuming that the background noise could fit a Gaussian distribution.

Finally, we measure the Sources to Artifacts Ratio (SAR) of each clipped file compared to its corresponding clean file using the Blind Source Separation Evaluation toolbox [16].

Clipping Rate(%)	Quality Measure			
	NIST STNR	WADA STNR	PESQ	SAR
0	49.6	80.7	4.1	—
0.5	48.8	83.8	3.9	17.9
1	48.2	83.1	3.7	15.7
5	44.6	77.7	3.1	10.4
10	40.8	71.5	2.7	8.3

Table 2: *Comparison of four speech quality measures for original clean speech, and different levels (0.5%-10.0%) of waveform clipping on TIMIT data.*

Table 2 presents results obtained when evaluating the data from all 630 speakers of the artificially clipped TIMIT dataset with the aforementioned speech quality measures. All four speech quality measures are clearly sensitive to the presence of clipping, with scores diminishing as the level of introduced clipping increases for each algorithm. We can see that the WADA SNR reported for the unclipped audio recordings is lower than both 0.5% and 1.0% clipped recordings, suggesting that the original TIMIT dataset may have some slight distortion. In addition, we can observe a clear degrading trend compared to the original unclipped audio. There is a high and consistent correlation between clipping rates and these quality measures, where 1% to 5% clipping having a much larger impact on speech quality scores compared with the distortion change from 0.5% to 1% clipping.

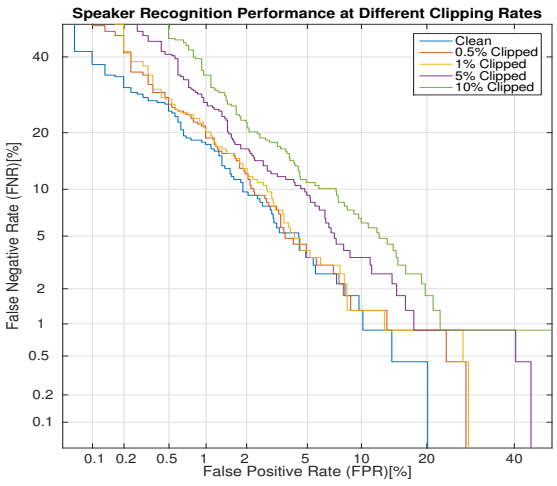

Figure 2: *DET curve results using i-Vector with PLDA approach on NIST SRE-08 corpus: performance for 0%, 0.5%, 1%, 5%, 10% clipping distortion.*

4.3. Impact of clipping for speaker recognition

Speaker recognition refers to the problem of recognizing a speaker from an unknown speech utterance given a pool of known target speakers. The system is usually trained on a large amount of speakers beforehand, and then we use a classifier to compare features of the enrolled target speaker with features extracted from the unknown test utterance [17]. We use the i-Vector approach [18, 19, 20] with PLDA-based scoring method [21, 22] on the NIST SRE-08 data to further investigate the effect of clipping.

4.3.1. I-Vector with PLDA

NIST SRE-08 data includes multiple microphone and recording environments such as conversational telephone speech and speech data recorded with a microphone in an interview scenario. Some speakers in the telephone conversational data are bilingual and their evaluation data may include non English speech as well. Here, we focus on the core short2-short3 test condition of the SRE08 data. This condition includes two-channel telephone conversational excerpts of about five-minutes total duration. We perform gender dependent speaker recognition on all male data. All training and test data involve only English language telephone speech spoken by native American English speakers. The speech data are sampled at 8kHz and coded in PCM format. We use all previous SRE data to train the background model and a total of 39,433 test trials are used in this evaluation.

We first extract 60 dimensional MFCCs, including delta and delta-delta acceleration vectors. A 2048 mixture Gaussian Background Model is trained on previous SRE data. We extract 400 dimensional i-Vectors $\boldsymbol{\omega}$ for each enrolled speaker and test utterance. Finally, we use the Probabilistic Linear Discriminant Analysis (PLDA) scoring method to perform the log-likelihood ratio test as shown in Eq. (1), where hypothesis $\mathcal{H}_s$ is that i-Vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are from the same speaker and hypothesis $\mathcal{H}_d$ assumes that they are generated from different speakers.

$$\text{score} = \log \frac{p(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2|\mathcal{H}_s)}{p(\boldsymbol{\omega}_1|\mathcal{H}_d)p(\boldsymbol{\omega}_2|\mathcal{H}_d)} \quad (1)$$

Fig. 2 shows the detection error trade-off (DET) curve of our speaker recognition system at 5 different clipping rates. From the figure and Table 3, we observe that with an increase in clipping rate, speaker verification performance drops correspondingly. When clipping rate increases from 5% to 10%, we see a significant performance loss with respect to Equal Error Rate (EER) and minimum Detection Cost Function (DCF). Due to the non-linear characteristic of speech signal clipping, the speaker dependent information contained in voiced segments are more likely to be clipped, if the input gain is set too high for a given recording. If only a very small amount of samples (e.g. less than 1%) are clipped, it may not have a notably negative affect for speaker recognition. This is because at 1% clipping rate, sufficient high energy voiced speech content which possesses significant speaker discriminant information remains intact.

Test Data	EER(%)	DCF
Clean	4.8	0.0257
0.5% clipped	4.4	0.0287
1% clipped	4.8	0.0289
5% clipped	6.6	0.0339
10% clipped	8.3	0.0408

Table 3: *EER and Minimum DCF results on NIST SRE-08 corpus: performance for clean trained speaker models and clean, 0.5%, 1%, 5%, 10% clipped test data.*

5. Conclusions and Future Work

In this study, we have investigated the causes and impact of the nonlinear distortion introduced by clipping. We explored clipping effects for speech quality based on objective speech quality assessment and a subjective human listening test. We also showed the impact of clipping on automatic speaker recognition systems.

Clipping has a degrading effect on the accuracy of speech systems including automatic speaker recognition. The effects range from a trivial drop in accuracy when only a small percent of data is affected, to a significant performance loss when clipping contamination is widespread. We find that if only a small number of files within a dataset suffer from clipping contamination, the negative performance impact will likewise be minimal. As the clipping rate increases to 5%, the negative impact that may reduce system performance is much higher. We should be aware of this phenomenon before processing speech data, and be responsible for inspecting whether the severity of clipping within a given data set is too high.

This study has therefore explored a number of issues where clipping contamination can impact speech perceptual quality and speaker recognition systems. CRSS will release the ClipDaT toolkit to the speech and language processing community to provide automatic detection to tag clipping occurrences, allowing for both speech corpus creators to detect and resolve issues early on. It is also important for researchers to be aware of the presence of clipping, and decide how to proceed with the affected speech data. Future work will consider robust speaker representations using generative adversarial networks in order to reduce the negative effect of clipping distortion for speech systems.

6. References

[1] S. Aleinik and Y. Matveev, "Detection of clipped fragments in speech signals," *Int. J. Electr. Electron. Sci. Eng*, vol. 8, no. 2, pp. 74–80, 2014, 00005.

[2] K. Simonchik, S. Aleinik, D. Ivanko, and G. Lavrentyeva, "Automatic preprocessing technique for detection of corrupted speech signal fragments for the purpose of speaker recognition," in *International Conference on Speech and Computer*, 2015, pp. 121–128.

[3] L. Ding, A. Radwan, M. S. El-Hennawey, and R. A. Goubran, "Measurement of the effects of temporal clipping on speech quality," *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 4, pp. 1197–1203, Aug. 2006.

[4] J. Eaton and P. A. Naylor, "Detection of clipping in coded speech signals," in *Signal Processing Conference (EUSIPCO), Proceedings of the 21st European*, 2013, pp. 1–5.

[5] J. Eaton and P. A. Naylor, "Noise-robust detection of peak-clipping in decoded speech," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2014, pp. 7019–7023.

[6] F. Bie, D. Wang, J. Wang, and T. F. Zheng, "Detection and reconstruction of clipped speech for speaker recognition," *Speech Communication*, vol. 72, pp. 218–231, Sep. 2015.

[7] F. Deng, C.-c. Bao, and F. Bao, "Clipping detection of audio signals based on kernel fisher discriminant," in *Signal and Information Processing (ChinaSIP), IEEE China Summit & International Conference on*, 2013, pp. 99–103.

[8] Y. Tachioka, T. Narita, and J. Ishii, "Speech recognition performance estimation for clipped speech based on objective measures," *Acoustical Science and Technology*, vol. 35, no. 6, pp. 324–326, 2014.

[9] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Monitoring the effects of temporal clipping on voip speech quality," in *INTERSPEECH*, 2013, pp. 1188–1192.

[10] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Monitoring voip speech quality for chopped and clipped speech," *Communications 2016*.

[11] S. Maymon, E. Marcheret, and V. Goel, "Restoration of clipped signals with application to speech recognition." in *INTERSPEECH*, 2013, pp. 3294–3297.

[12] M. J. Harvilla and R. M. Stern, "Least squares signal declipping for robust speech recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[13] M. J. Harvilla and R. M. Stern, "Efficient audio declipping using regularized least squares," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2015, pp. 221–225.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on*, vol. 2, 2001, pp. 749–752.

[15] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis." in *INTERSPEECH*, 2008, pp. 2598–2601.

[16] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[17] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[18] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[19] P. Matějka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Černocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2011, pp. 4828–4831.

[20] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[21] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[22] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision (ICCV), IEEE International Conference on*, 2007, pp. 1–8.