# Project Notes:

**Project Title:** Minimizing Language Dependency in a Voice Authentication System
**Name:** Ryan Mechery

**Note Well:** There are NO SHORT-cuts to reading journal articles and taking notes from them. Comprehension is paramount. You will most likely need to read it several times so set aside enough time in your schedule.

**Contents:**

# Knowledge Gaps:

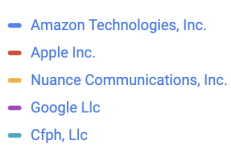
This list provides a brief overview of the major knowledge gaps for this project, how they were resolved and where to find the information.

| Knowledge Gap | Resolved By | Information is located | Date resolved |
|---|---|---|---|
| What is MFCC? | • Reading an article directly on the topic.<br>• Doing my own research into signal processing. | • Article #3 | 8/29/2021 |
| Is machine learning the best solution for voice authentication? | • Reading articles on the difference between text-dependent and text-independent speaker authentication systems. | • Article #5 | 9/12/2021 |
| What features are extracted by speaker verification systems? | • A list (in-progress) can be found here. Compiled by searching features from articles but understood them more fully by watching a series on audio signal processing. | • Almost all articles<br>• This website had a neat list (added in some more).<br>• Audio Signal Processing for Machine Learning by Valerio Velardo. | 10/21/2021 |
| How are those features extracted? | • Searched online to find a list of various feature extraction methods and learned about each one in detail through most of the Articles | • Article #3<br>• Article #4<br>• Article #5<br>• Article #7<br>• Article #8<br>• Article #9<br>• Article #10 | 10/7/2021 |
| Why do different languages used in training affect Speaker Verification systems? | • Read about an introduction to phonetics and how language on a signal processing level. | • Article #11 | 10/17/2021 |
| What is an SVM? | • Read the scikit documentation to get a definition and analysis. | • sci-kit<br>• Article #16 | 11/23/2021 |

| | | | |
|---|---|---|---|
| How are SVMs used in Speaker Verification? | • Read a journal article with that implementation. | • Article #16 | 11/27/2021 |

# Literature Search Parameters:

These searches were performed between Aug 30, 2021 and Nov 24, 2021.
List of keywords and databases used during this project.

| Database/search engine | Keywords | Summary of search |
|---|---|---|
| Google Scholar | Biometrics | Biometrics are an authentication method that uses unique facial features to validate a user. There are many types like fingerprint, iris, and voice recognition. |
| Google Scholar | “Voice Biometrics” AND “Voice Authentication” | Voice biometrics are a method of authentication that uses your voice to validate a user. They can either be text-dependent or text-independent. Voice authentication is more vulnerable to spoofing attacks. Voice authentication can be used over the phone and they don’t need specific hardware compared to fingerprinting and iris recognition. |
| Google Scholar | ("text-dependent" AND "text independent") AND "speaker verification" | Just one article appeared in the search. That was Article #5. |
| USPTO Patent Full-Text and Image Database (PatFT) | Voice Authentication OR Voice Biometric [all fields] | Many results came up. Many had some variation of “systems and methods for voice authentication systems” and others didn’t seem to be related to voice biometrics. |
| Google Scholar | speaker recognition tutorial | Most results were fairly old and were written more as review articles rather than in-depth tutorials. |
| Google Patents | Speaker verification | Similar results to the USPTO search. Image to the right shows assignees of patents.   |
| Google Scholar | "speaker verification" AND | These results came up. I have come |

| | | |
|---|---|---|
| | "languages" | across some of these papers before, but searching with these specific keywords have allowed me to find even more articles on SV systems in relation to language dependency. |
| Google Scholar | language dependency AND ( voice authentication OR speaker verification OR speaker recognition) | Many papers I have already come across appeared in this search. However this specific search paramater allowed me to find a couple more papers that deal with speaker verification in relation to language dependency. |
| Google Scholar | speaker verification AND svm | I chose the “SVM” keyword to look at SV systems using Support Vector Machines as a possible |

# Article #0 Notes: Title (EXAMPLE)

Article notes should be on separate sheets

**KEEP THIS BLANK AND USE AS A TEMPLATE**

| | |
|---|---|
| Source Title | |
| Source citation (APA Format) | |
| Original URL | |
| Source type | |
| Keywords | |
| Summary of key points + notes (include methodology) | |
| Research Question/Problem/ Need | |
| Important Figures | |
| VOCAB: (w/definition) | |
| Cited references to follow up on | |
| Follow up Questions | |

# Article #1 Notes: Study on a Biometric Authentication Model based on ECG using a Fuzzy Neural Network

| | |
|---|---|
| Source Title | Study on a Biometric Authentication Model based on ECG using a Fuzzy Neural Network |
| Source citation (APA Format) | Kim, H. J., & Lim, J. S. (2018). *Study on a Biometric Authentication Model based on ECG using a Fuzzy Neural Network*. *317*, 012030. https://doi.org/10.1088/1757-899x/317/1/012030 |
| Original URL | https://iopscience.iop.org/article/10.1088/1757-899X/317/1/012030/pdf |
| Source type | Conference Paper |
| Keywords | True Acceptance Rate (TAR) - Probability of an authentication system correctly validating a user.<br>False Acceptance Rate (FAR) - Probability of an authentication system incorrectly validating a user.<br>False Reject Rate (FRR) - Probability of an authentication system incorrectly rejecting a validated user. |
| Summary of key points (include methodology) | Researchers at Gachon University have proposed a faster and more accurate model that uses ECG readings, which are recordings of the heart's electrical signals, as a Biometric Authentication Method. The authentication uses something known as fuzzy neural networks which require previous expert knowledge but are much simpler to understand and use. The researchers tested their proposed model on 73 participants from the Physionet Database and on average, the true accept rate (TAR) for authentication was 98.32% and the false accept rate (FAR) was 5.84%. |
| Research Question/Problem/ Need | Problem: Current ECG biometric systems are slow and not accurate enough. |
| Important Figures | |

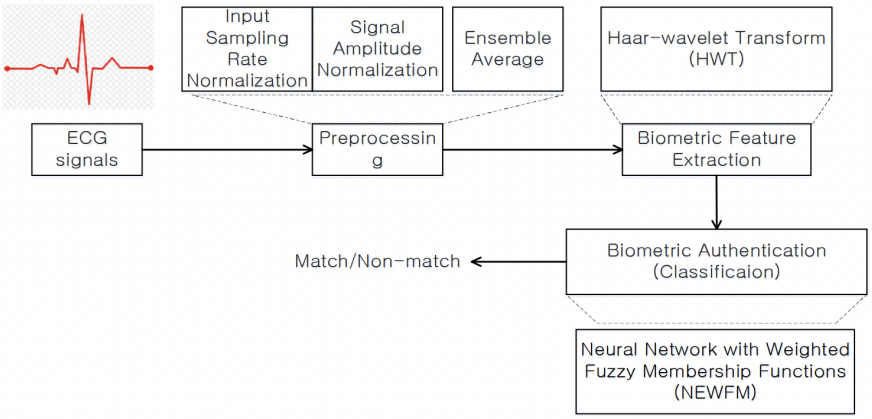



**Figure 1** is a flowchart that shows a model of the ECG biometric authentication method.

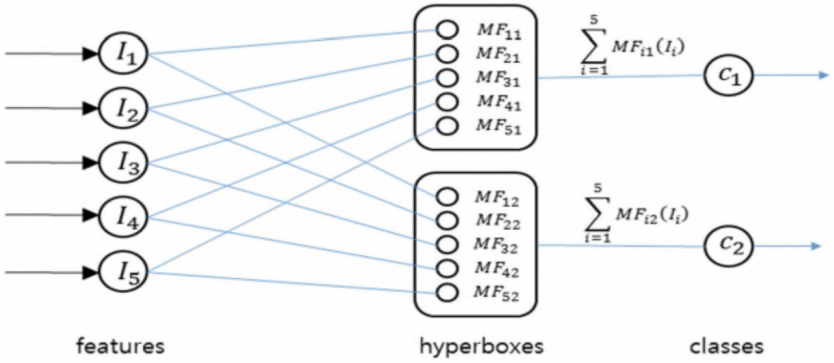



**Figure 5** is a diagram that shows the design of the fuzzy neural network.

| Algorithm | Subjects | TAR | FAR | Authent. lenght | Enroll lenght |
|---|---|---|---|---|---|
| Singh et al.[26] | 73 | 82.00% | 7.00% | 1/2 of record | 1/2 of record |
| Wubbeler et al.[8] | 74 | 97.00% | 3.00% | 10 s. | 10 s. |
| singh et al. [30] | 73 | 95.55% | 3.00% | 10 heart beat | 10 heart beat |
| Juan et al.[28] | 73 | 84.93% | 1.29% | 4 s. | 30 .s |
| Proposed Algorithm | 73 | 98.32% | 5.84% | 1 heart beat | 15 heart beat |

**Table 1** lists the results of the testing compared to other ECG biometric systems.

Notes

Follow up Questions

1) How does using this ECG biometric compare to other authentication methods?
2) Would the biometric authentication method fail in the event of abnormal heart activity?
3) Would the proposed model be more accurate if the ECG recording time increases?

# Article #2 Notes: Experimental Feeding Regime Influences Urban Bird Disease Dynamics

| | |
|---|---|
| Source Title | Experimental feeding regime influences urban bird disease dynamics |
| Source citation (APA Format) | Galbraith, J.A., Stanley, M.C., Jones, D.N. and Beggs, J.R. (2017), Experimental feeding regime influences urban bird disease dynamics. J Avian Biol, 48: 700-713. https://doi.org/10.1111/jav.01076 |
| Original URL | https://onlinelibrary.wiley.com/doi/abs/10.1111/jav.01076 |
| Source type | Research Paper (Abstract) |
| Keywords | n/a |
| Summary of key points (include methodology) | Researchers in New Zealand have studied how feeding wild birds is directly correlated to avian disease epidemics. Over a year and half, the researchers studied both fed, and non-fed, wild birds to examine their health condition and see if they had any pathogens. Researchers found birds had a 7% chance of testing positive for Salmonella Typhimurium and species like the *Zosterops lateralis,* declined in population due to their feeding practices. |
| Research Question/Problem/ Need | Question: Is feeding wild birds directly correlated to transmission of avian diseases? |
| Important Figures | None (Just an Abstract) |
| Notes | none |
| Cited references to follow up on | n/a |
| Follow up Questions | 1) Are bird feeders, or the bird feed itself spreading the pathogens?<br>2) Would birds increase in population if they were stopped being fed?<br>3) Are birds in urban areas more likely to spread pathogens compared to rural areas? |

# Article #3 Notes: MFCC and its applications in speaker recognition

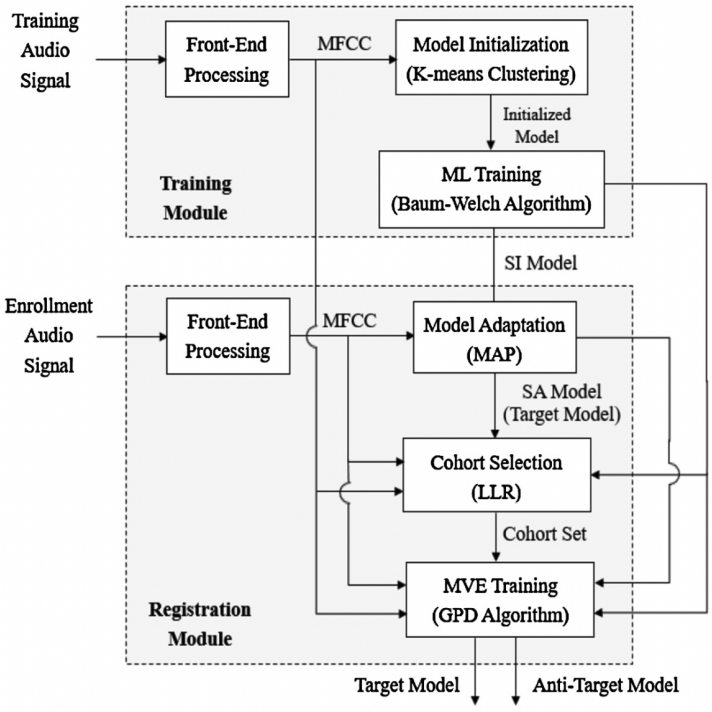
| | |
|---|---|
| Source Title | MFCC and its applications in speaker recognition |
| Source citation (APA Format) | Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International journal on emerging technologies*, *1*(1), 19-22. |
| Original URL | http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.689.4627&rep=rep1&type=pdf |
| Source type | Journal Article |
| Keywords | Feature extraction - process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal. (https://www.ee.iitb.ac.in/~esgroup/es_mtech03_sem/sem03_paper_03307003.pdf)<br>Mel frequency cepstral coefficients (MFCC) - feature extraction technique based on the mel scale that involves windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale (https://link.springer.com/content/pdf/bbm%3A978-3-319-49220-9%2F1.pdf)<br>Speaker recognition - Speaker verification is the process of accepting or rejecting the identity claimed by a speaker. (Scholarpedia)<br>Window - a function (shape) that is nonzero for some period of time, and zero before and after that period (UCI) |
| Summary of key points (include methodology) | Vibha Tiwari, a researcher at the Gyan Ganga Institute of Technology and Management, wrote about the field of digital signal processing in this article and their goal was to create a text-dependent speaker recognition system. Since each person's voice tract is anatomically different, it can be used as a reliable biometric which has the advantage of being able to be used for remote authentication. "Like any other pattern recognition systems, speaker recognition systems also involve two phases namely, training and testing.(Tiwari, 2010, p1)" During the training process, the first step is to extract features from an audio sample which can then be used to authenticate a user later down the line. Feature extraction involves extracting distinct features from audio such as pitch, and sending it to a pattern classifier which categorizes the given data. Some techniques for feature extraction are Mel-frequency cepstral coefficients (MFCC), Linear predictive coding (LPC), and Local discriminant bases (LDB). |

<table>
<tr><td></td><td>“MFCC is based on the human peripheral auditory system (Tiwari, 2010, p2),“ and it is a direct relation between perceived frequency, and measured frequency of a tone, which can be displayed on a graph and stored in a vector. MFCC calculation has two steps: Mel-frequency wrapping, which converts frequency into the mel scale, and Cepstrum, which is where the mel scale is converted to time. In their approach, they used multiple filters to amplify certain frequency ranges and they also used different windows, or functions to plot the data from the MFCC. In their research, Tiwari compared different implementations of MFCC. They found that MFCC with 32 filters resulted in an efficiency of 85%, which was the highest compared to 12, 22, and even 42 filters. They also found out that using a Hanning window had a 20% better overall efficiency at 75% compared to using a Rectangular window at 55% efficiency. In order to improve the system, they came to the conclusion that the speaker recognition system had to be built without being text-dependent to offer greater efficiency.</td></tr>
<tr><td>Research Question/Problem/ Need</td><td>Engineering Need: Can modifications be made to the existing technique of MFCC for feature extraction?</td></tr>
<tr><td>Important Figures</td><td>Figures weren’t labelled in this journal article but I found the concluding table important because it compared the efficiency of all the parameters making it easy to understand the results.
<table>
<tr><td>Number of filters</td><td>12</td><td>22</td><td>32</td><td>42</td></tr>
<tr><td>Efficiency</td><td>65%</td><td>75%</td><td>85%</td><td>80%</td></tr>
</table>
<table>
<tr><th>Types of window using 32 filters</th><th>Efficiency</th></tr>
<tr><td>Hanning</td><td>75%</td></tr>
<tr><td>Rectangular</td><td>55%</td></tr>
</table></td></tr>
<tr><td>Notes</td><td></td></tr>
<tr><td>Cited references to follow up on</td><td>None. Many of the articles are very outdated and aren’t exactly relevant.</td></tr>
<tr><td>Follow up Questions</td><td>1) Would the proposed system have the same efficiency if it was text-independent compared to its current text-dependent status?<br>2) Why did having more filters decrease the efficiency of the speaker recognition system?</td></tr>
</table>

# Article #4 Notes: Active Voice Authentication

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Active Voice Authentication |
| Source citation (APA Format) | Meng, Z., Altaf, M. U. B., & Juang, B.-H. (Fred). (2020). Active voice authentication. Digital Signal Processing, 101, 102672. https://doi.org/10.1016/j.dsp.2020.102672 |
| Original URL | https://www.sciencedirect.com/science/article/abs/pii/S1051200420300178 |
| Source type | Peer-Reviewed Journal Article |
| Keywords | 1) Voice Activity Detector (VAD)<br>2) NIST Speaker Recognition Evaluation (SRE) - Dataset used to evaluate accuracy of speaker recognition systems.<br>3) Maximum a Posteriori (MAP) adaptation - acoustic model adaptation technique in speech recognition. Most common.<br>4) Minimum Verification Error (MVE)<br>5) Hidden Markov Model (HMM) - statistical method for modeling sequential data<br>6) Likelihood Ratio Test<br>7) Gaussian Mixture Models (GMMs) - statistical method for grouping data points that to a single distribution<br>8) Factor Analysis<br>9) Speech Frame - Length of signal segment.<br>10) Window - Test segment of a signal. Expressed in a number of frames.<br>11) I-Vector Analysis |
| Summary of key points (include methodology) | The engineering need that this paper focused on was improvements that can be made Active Voice Authentication (AVA) or text-independent voice authentication. The purpose of this study was to design a novel AVA framework, propose a window based testing scheme that allows for real-time testing, apply MAP adaptation to minimize enrollment data, and apply Minimum Verification Error (MVE) training to minimize speaker verification error. In order to get test data, the researchers created an AVA database from 25 speakers who, each, provided 2.5 hours of speaking data in the form of sentences and numbers. In experimentation, they first used a tool for conventional voice authentication, called i-vector analysis, to see if VA method was adequate for voice authentication. Next, they built their AVA system which contained 3 parts, a training module, a registration module, and an authentication module. Using the test |

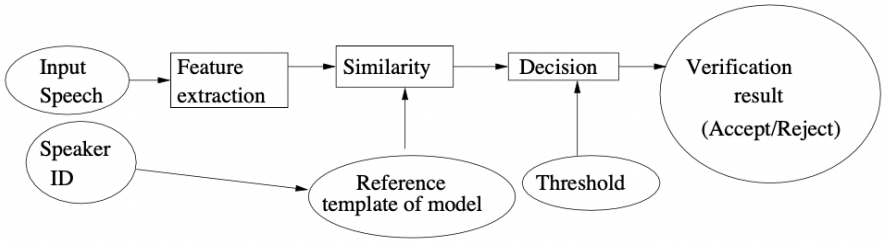
| | |
|---|---|
| | data, the researchers used machine learning to create a MAP adapted model. This model was put through another round of training called MVE training to limit verification errors. To ensure that a hacker couldn't use spoofing methods to use a digital voice, they added a Voice Activity Detector (VAD) into the design to ensure that a person was present and speaking for every frame. They used a statistic called Window Based Estimation Rate (WEER) to measure results. Through experimentation, the researchers made a number of findings. The first was that Conventional VA wasn't accurate enough to be used as a biometric. Although they used many hours of test data, researchers found that only 180 seconds of voice data was needed at enrollment as it offered a perfect tradeoff between accuracy and performance in VA. Third, using 180 seconds of voice data offered 3-4% WEER while 6-10 minutes of enrollment data was needed to raise that percentage to 0.5 to 1% WEER. And last was that using an external microphone that was different from the one used in training would require a rebuilding of the AVA model. |
| Research Question/Problem/ Need | Problem: Can a novel Active Voice Authentication system be made that continuously tests input data and provides a greater accuracy than CVA? |
| Important Figures |  <br>**Figure 5** shows a flowchart of the AVA design. |
| Notes | • **Conventional Voice Authentication** is also known as text-dependent speech authentication where an audio clip is |

| | |
|---|---|
| | *stored* at enrollment and it is directly compared at verification.<br>○ Requires the user to recite a vocal passphrase or digits.<br>○ This is very insecure as a hacker could easily spoof the verification system using an audio sample of the user's voice.<br>○ SVA only analyzes the speaker sample *after* the verification process.<br>● **Active Voice Authentication (AVA)** is also known as text-independent speech authentication where an audio clip is *analyzed* for its vocal features such as pitch and tone, which are then compared to the features of an audio sample at verification.<br>○ A system with AVA might still use audio samples of a user saying text to build a model. However, the AVA system will analyze the user's voice, not simply store the sample to be compared later.<br>○ This is much more secure as hackers would need to create an AI to accurately reproduce your voice in order to spoof the system.<br>○ If proper training is done and enough features are extracted during the enrollment and training phase, then the system can validate a user with a very short audio sample.<br>○ AVA is *continuously* analyzing the speaker sample and denies access to users that take an excessive amount of time. |
| Cited references to follow up on | 1) J. Markel, B. Oshika, A. Gray Jr., Long-term feature averaging for speaker recog- nition, IEEE Trans. Acoust. Speech Signal Process. 25 (4) (Aug 1977) 330–337.<br>2) A.E. Rosenberg, Olivier Siohan, S. Parathasarathy, Speaker verification using minimum verification error training, in: 1998 IEEE ICASSP, vol. 1, May 1998, pp. 105–108. |
| Follow up Questions | 1) Is a window duration longer than 1.01 seconds possible with modern computing advancements?<br>2) Would the model perform with the same level of accuracy if the user's voice changes? (E.x. User gets a sore throat and has a raspy voice.)<br>3) Although researchers had noted that spoofing would be ineffective with AVA, would the novel model be able to recognize digital sounds trying to sound like a human? |

# Article #5 Notes: Text dependent and text independent speaker verification system:

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Text dependent and text independent speaker verification system: Technology and application. |
| Source citation (APA Format) | Marinov, S. (2003). Text dependent and text independent speaker verification system: Technology and application. *Overview article*. |
| Original URL | https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.1529&rep=rep1&type=pdf |
| Source type | Overview Article |
| Keywords | Biometrics - the statistical study of biological phenomena (dictionary definition)<br>Speaker Verification (SV) - the process of verifying the claimed identity of a registered speaker by using their voice characteristics. Note: Not to be confused with Speaker Recognition (SR). SV is *apart* of SR.<br>Spectral Analysis - calculation of waves or oscillations in a set of sequenced data. (ScienceDirect)<br>Utterance - a spoken word, statement, or vocal sound. Emphasis on *spoken word* in speaker verification systems. (Oxford Dictionary) |
| Summary of key points (include methodology) | The research question that this paper aimed to address was, “What are the differences between Text Dependent and Independent speaker verification?” Since this was an overview article, no experiment was conducted, but the goal of the paper was to lay out the differences, and list the applications of both technologies. In the introduction, Marinov explained the different types of authentication methods, and said that voice biometrics are a type of behavioral biometric because they are prone to change. Marinov then explained the two parts of a speaker verification model, the enrollment module and the verification module. The enrollment module is used to build a model of a user initially, and the verification module collects audio data from a user at startup, extracts features from it, and compares it to a stored model. Based on those results, the verification module gives a result of “Accept” or “Reject.” Spectral analysis is the first step in the enrollment phase and it is where segments of an audio sample are analyzed with signal processing feature extraction methods like LPC and MFCC. Normalization techniques are statistical techniques that aim to mitigate the variability that can occur |

| | |
|---|---|
| | from different audio samples (e.x. User gives an audio sample in the training phase at home but gives an audio sample in the verification phase outside). These are components all SV systems should have but Text Dependent SV (TDSV) systems and Text Independent SV (TISV) systems are very different. TDSV systems rely on a predefined utterance in training the system. This means that TDSV systems cannot differentiate between validated users and impersonators with audio recordings of validated users because the system is purely relying on *what* the user is saying. On the other hand, TISV systems analyze audio recordings in much more detail to focus on *how* a user is saying something (e.x. Tone and pitch). This system requires machine learning as the system needs to extract vocal features that cannot be directly obtained from the voice data. To state again, this paper didn't conduct an experiment, but the author still discovered some findings from his data. First is that although users can say whatever they want in a TISV system, the system will be more accurate if the user speaks in the same language. The second finding is that more research needs to be done in the applications of SV systems to determine where they are the most effective. |
| Research Question/Problem/ Need | Question: What are the similarities and differences between Text Dependent and Text Independent speaker verification systems? |
| Important Figures |  <br>**Figure 1** is a flowchart that shows the decision making process of a SV system. |

| Characteristics | Fingerprints | Hand Geometry | Retina | Iris | Face | Signature | Voice |
|---|---|---|---|---|---|---|---|
| Ease of Use | High | High | Low | Medium | Medium | High | High |
| Error incidence | Dryness, dirt, age | Hand injury, age | Glasses | Poor lighting | Lighting, glasses, age, hair | Changing signature | Noises, colds, weather |
| Accuracy | High | High | Very High | Very High | High | High | High |
| User Acceptance | Medium | Medium | Medium | Medium | Medium | Medium | High |
| Required Security Level | High | Medium | High | Very High | Medium | Medium | Medium |
| Long-term stability | High | Medium | High | High | Medium | Medium | Medium |

This is an **unlabelled table** that gives comparisons of different biometrics. This table is quite useful as it conveys the security of voice biometrics compared to other biometrics. Although Voice Biometrics has high ease of use, it is variable in noisy conditions and doesn't have a high long-term stability.

Notes

- Authentication Types:
  - something you know - a password, PIN, or piece of personal information;
  - something you have - a card key, smart card, or token;
  - something you are - a biometric.
    - Most secure one but once it's stolen, you can never reset this type of password.
- Types of Biometrics:
  - Physical Biometrics - "Impossible" to change. Possible to be scanned without your knowledge.
    - Fingerprints
    - hand or palm geometry
    - Retina
    - Iris
    - facial characteristics
  - Behavioral Biometrics - Prone to change. Cannot directly be scanned, needs machine learning since it

is dependent on feature extraction.
    - Signature
    - Voice
    - keystroke pattern
    - Gait
- Process of SV
    1) **Enrollment** - User supplies voice data and then a model is built using it. AKA Training Period.
        - Getting training data from new user
        - Spectral Analysis
        - Training model
    2) **Verification** - Compares data from data current audio sample to previously stored results and gives a binary result of acceptance or rejection.
        - Utterance
        - Spectral Analysis
        - Comparison
        - Decision (Acceptance/Rejection)
- Spectral Analysis
    - First step in the enrollment process and involves feature extraction.
    - Essentially, you are analyzing waves in sequenced audio data.
        - For SV, you look at short audio segments like 20ms.
    - Types of Spectral Measurements
        - Linear Predictive Coding (LPC)
        - Mel Frequency Cepstral Coefficients (MFCC)
- Normalization Techniques
    - Many factors cause your voice to change making it hard to compare audio samples of a user at different times.
    - So, normalization techniques are used to minimize variations that interfere with the verification process.
    - Techniques:
        - **Blind Equalization Method**: Effective for text dependent SV and works optimally if a long utterance, or audio sample of a user speaking, is used. Essentially, “cepstral coefficients are averaged over its duration and then these values are subtracted from the cepstral coefficients of each frame. (Marinov, p.5)”
        - **Probability Method**: a likelihood ratio dependent is calculated based on 2 probabilities: the likelihood that the verification audio is from a validated user, and the likelihood that the audio is from an

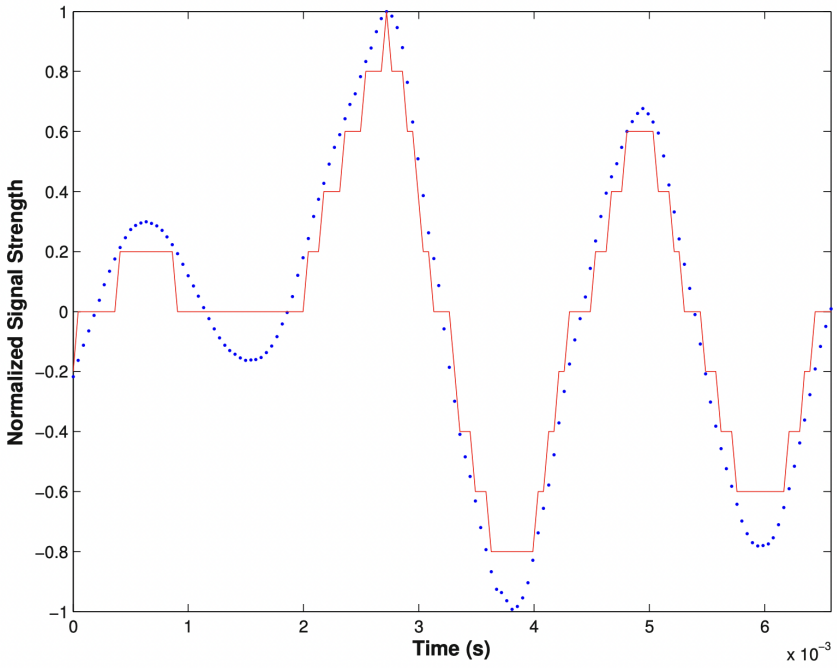
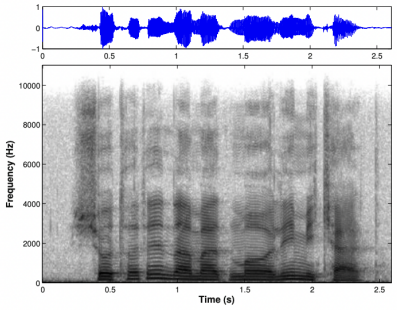
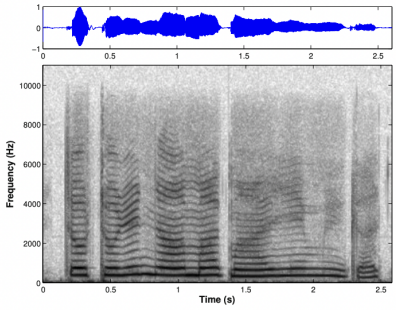
| | |
|---|---|
| | impersonator.<br>● Text Dependent Speaker Verification (TDSV)<br>○ Predefined utterance is used in training the system.<br>■ This means TDSV cannot distinguish between a validated user, and an impersonator that stole audio recordings from the validated user.<br>○ Examples of TDSV systems are where a user says all digits at enrollment and a passphrase with a random number is given at verification.<br>● Text Independent Speaker Verification (TISV)<br>○ Authentication isn't dependent on *what* the user is saying, but rather *how* they are speaking. This means that users aren't limited to saying passphrases or numbers.<br>○ Machine Learning is needed in this step.<br>○ Methods:<br>■ Long-term statistics and multi dimensional autoregressive<br>■ Vector Quantization<br>■ Fully Connected Hidden Markov Models (HMM)<br>■ Artificial Neural Network<br>■ Gaussian Mixture Models (GMM)<br>○ Although users can say whatever, TISV systems are more accurate when users talk in the same language.<br>● Although SV systems only give two results, Accept or Reject, these systems actually have 4 results: True Accept, False Accept, True Reject, False Reject.<br>○ This is because there is always a chance that the system can make the wrong decision. This is why there are four main statistics used to test the accuracy of biometric systems: TAR, FAR, TRE, FRE.<br>● Knowledge Gap #2: This article, and others, have shown that TISV systems are much more accurate than TDSV systems. So, in order to create a reliable TISV system, machine learning is needed because feature extraction |
| Cited references to follow up on | 1) Blomberg, Mats. 2002. Speaker Verification. Slides from Introductory Lectures.<br>(Note: Author referenced this source many times)<br>2) Mathew, M., B. Yegnanarayana, and R. Sundar. 1999. A neural network-based text-dependent speaker verification system using suprasegmental features. url= citeseer.nj.nec.com/404364.html.<br>3) Rydin, Sara. 2001. Text dependent and text independent speaker verification systems. technology and applications. Term paper in Speech Technology. |

| | |
|---|---|
| Follow up Questions | 1) If a TISV system depends on feature extraction of vocal characteristics, then why is the system more accurate if the user speaks the same language?<br>2) Would a SV system have greater accuracy if the segment length is less than the estimated 20ms?<br>3) Why does Voice Biometrics have a rating of “Medium” for Long Term Stability when a user’s voice can change drastically throughout the course of their life; anywhere from 10 to just 2 years? |

# Article #6 Notes: Signal Representation of Speech (Chapter 3)

| | |
|---|---|
| Source Title | Signal Representation of Speech (Chapter 3) |
| Source citation (APA Format) | Beigi, H. (2011). *Fundamentals of Speaker Recognition*. Springer US. https://doi.org/10.1007/978-0-387-77592-0 |
| Original URL | https://link-springer-com.ezpv7-web-p-u01.wpi.edu/content/pdf/10.1007%2F978-0-387-77592-0.pdf |
| Source type | Book Chapter |
| Keywords | Signal - An observed measurement of a physical phenomenon. It generally describes an observation of a higher level physical phenomenon in correlation with lower level measurement concepts such as time or space.<br>Sampling - Process to convert a signal from continuous time to discrete time<br>Non-Stationary Signal - A signal whose statistical parameters change over time. Parameters such as intensity, variance, etc.<br>Quantization - the process of mapping continuous infinite values to a smaller set of discrete finite values<br>Spectrogram - Graph that shows signal strength of a signal over time at various frequencies. (pnsn.org)<br>Formants - Resonant (Clear) Regions of a spectogram |
| Summary of key points (include methodology) | Fundamentals of Speaker Recognition is a full-length publication that is an overview of all the technologies used in speaker verification. I read one chapter of the book entitled, “Chapter 3: Signal Representation of Speech.” This chapter covered signal processing methods and techniques that are applicable to Speaker Verification (SV) systems. In the introduction, Beigi first defined what a signal was; it is just a measurement of some kind of information (in SV systems it’s audio signals). There are two types of signals: analog and discrete. *Analog signa*ls can be represented with a mathematical function meaning that you can find an exact dependent variable with a known variable such as t. *Discrete signals* are a sequence of points and they are mapped from an analog signal in a process known as *sampling*. *Quantization* is another conversion technique that aims to digitize the amplitude value, whereas sampling digitizes the coordinate value. It is dependent on bitrate and is expressed by dividing amplitude by the number of values. Computers process |

| | |
|---|---|
| | audio data with binary numbers, but visual representations make it easier to better understand the data. One type of audio visualization is a speech waveform (similar to what you would see in Audacity) which shows oscillation of sound and plots amplitude of signal for each time value. Another type is a spectogram which shows audio data in 3 dimensions (time, frequency, and signal strength/energy level). When you analyze a spectrogram of a person speaking, you can notice waves of lines. They are called *formants* and they are indirectly related to vocal tract length. This means that adult males have shorter formants compared to women and children. Since you cannot perfectly sample an analog signal, some sampling errors will occur. One type is aliasing and it's distortion or noise that comes from sampling. You can fix this error with a technique called anti-aliasing which lowers the quality or bitrate of the signal to normalize the extraneous noise. Another type of sampling error is loss of information and it's governed by the fact that you lose more and more information when you lower the sampling frequency. This observation can be seen in Figure 3.3 where the discrete signal has lost a lot of information at certain t values from the analog signal. Although no experiment was conducted, Beigi noted some important findings that were relevant to speaker verification systems. The first is that speaker verification is done using discrete signal processing because analog signal processing is too complex and hardware intensive to work with. The second is that although there are many sampling techniques, Speaker Verification (SV) systems most often use *periodic* sampling. The third finding is that quantization becomes very important in SV systems because trying to store quantized values of high bit rate values increases storage size, meaning that it needs more computing power to analyze. The fourth finding is that quantization is also important because SV systems are aimed at a variety of devices, which all use different types of microphones, and use different bitrates. Thus, lower bitrate and noise-prone microphones will impact accuracy of the system because data is already lost and distorted before signal pre-processing. The fifth finding is that formants become important because they are important parts of intelligible speech, but they are different in ages in sexes meaning that the same system will not have the same accuracy in authentication across a variety of demographics not directly related to sound. The fifth finding is that although truncation errors are a sampling error a programmer has to consider in their design with signal processing, it is **not** important to try and fix in SV systems because you don't need to reconstruct sampled data back to a function. |
| Research Question/Problem/ | RQ: Why is signal representation of speech important in Speaker Verification systems? |

| Need | |
|---|---|
| Important Figures |  <br>**Figure 3.3** shows how sampling results in loss of data using an analog signal that has been quantized 11 times.<br>  <br>**Fig. 3.18:** Adult male (44 years old)<br>  <br>**Fig. 3.19:** Male child (2 years old)<br>**Figure 3.18 and 3.19** compare two different spectrograms to show how age impacts formants. Adult formant is short and straight while child's formant is long and wavy. |

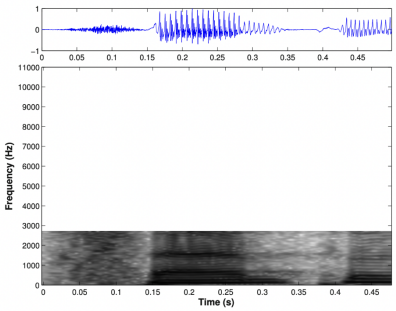



**Fig. 3.26:** Original signal was subsampled by a factor of 4 with no filtering done on the signal

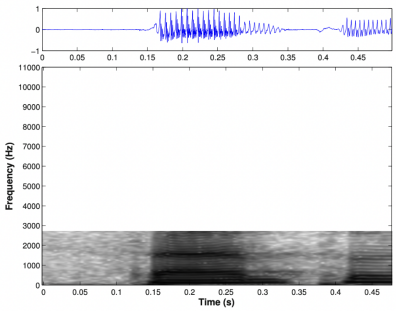


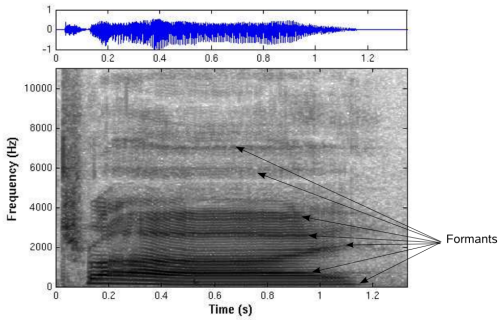

**Fig. 3.27:** The original signal was subsampled by a factor of 4 after being passed through a low-pass filter

**Figure 3.26 and 3.27** show how Anti-Aliasing techniques reduce noise and allow most of the speech to be extracted from original audio. In Fig. 3.27, you can see how noise from t=0s to t=0.15s was removed when the original signal from Fig. 3.26 was passed through a low-pass filter.

Notes

- **Introduction**
  - A **signal** is just a measurement or function of some information like audio, image, or electrical signals.
  - It's a mapping of a point from space or time into a higher level of measurement.
  - Continuous signals work in an analog domain because they are continuously moving through time and space.
  - A speech signal is important in speaker recognition.
  - Since you can technically take an infinite amount of information in a certain domain, in a process called **sampling,** you map a function into a sequence of points called a **discrete signal**.
  - **Imp**: Speech is a non-stationary signal.
    - Because of the way humans speak, the speech signals we produce have constantly changing parameters.
  - **Imp**: Speaker recognition is really only done through discrete signal processing because analog signal processing is too complex and hardware intensive to work with.
- **Sampling Audio**
  - Speech recognition is a passive process because you analyze the audio signal after it is spoken.
    - Doesn't mean you can't continuously check audio signals.
  - When audio is inputted through a microphone, it has to be converted to a digital signal so that you can

actually process.
    - Various sampling techniques include: periodic, cyclic, multirate, random, and pulse width modulated sampling.
    - **IMP:** Most speaker recognition systems use **periodic sampling** with techniques like PCM.
- **Quantization and Amplitude Errors**
    - Quantization is the process of converting infinite values from an analog signal to a finite set of values for a discrete signal.
    - Figure 3 shows the impact of quantizing an analog signal 11 times. Data is accurate in some places but it is pretty far off for other t values.
    - **Quantization Level** is dependent on bit rate of signal. 8 bit audio only has 256 values whereas 16 bit audio has 65,536 values. Calculated by dividing amp by number of values.
        - **IMP**: Quantization level changes storage size.
    - **IMP**: Quantization becomes important in Speaker Recognition because users could have different microphones. If they have a cheap microphone with low bit rate audio, that will affect the accuracy of the system.
    - You can try to use a filter to try and filter out things like ambient noise, but that distorts the original signal.
- **Speech Waveform**
    - Using amplitude quantization will allow you to get a speech signal which you can graph into a speech waveform. (Looks like what you would see in programs like Audacity)
- **Spectrogram**
    - A spectrogram is a 3 dimensional representation of spectral content of speech signal.
    - Shows time on x-axis, frequency on y-axis, and energy level for each frequency component through a shade of color. (White represents 0 energy and black is highest amount of energy)
    - **IMP**: Will need to look at this graph in Jupyter Notebooks so I need to research how to do that.
- **Formant Representation**
    - Formants are clear regions on a spectrogram. It's hard to explain but a clear formant will have straight lines while a less clear formant will have curvy/blurry lines.




- **IMP**: The longer the vocal tract, the shorter the formant. As expected, the shorter the vocal tract, the longer the formant. This is important because anatomically speaking, Adult Males have longer vocal tracts compared to females and children.
  - Variability from male to female formant length is 20%.
- Since formants are all about speech clarity, an adult will have a much clearer voice and shorter formant when they speak compared to a child.
- **IMP**: Formants are essential components in intelligible speech. (Speech with words)

Sampling Errors: You cannot perfectly sample an analog signal to a discrete signal.

- **Aliasing**
  - Distortion that comes about when a waveform is reconstructed from samples.
  - In signal processing, Anti-Aliasing is a technique used to minimize distortion from aliasing effects by passing audio through a low pass filter which decreases frequency of sampling rate.
- **Truncation Error**
  - In math, it's the derivative of a function <u>minus</u> the numerical approximation of the derivative.
  - Technique that allows for equations from sampling theorems to be computed.
  - **IMP**: <u>Not important</u> in speaker verification systems since you don't need to reconstruct the sample.
- **Loss of Information**
  - Essentially, you lose more and more information contained in an analog speech signal when you use a low sampling frequency.
  - The aim is to use the highest sampling frequency possible.
    - Time, Cost, and Computational factors are variables that must be considered in the final design of the Speaker Verification system.

| | |
|---|---|
| Cited references to follow up on | Many were on fundamental theorems in signal processing which isn't really relevant to my project. |
| Follow up Questions | 1) How formant lengths affect accuracy of SV systems?<br>2) How do different microphones affect accuracy of SV systems?<br>3) At what point does lower quantization level of data provide diminishing returns? (Is there a sweet spot for quantization that is good in time, computing power and storage size for speaker verification) |

# Article #7 Notes: Language Dependency In Text-Independent Speaker Verification

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Language Dependency In Text-Independent Speaker Verification |
| Source citation (APA Format) | Auckenthaler, Roland, Michael J. Carey, and John SD Mason. "Language dependency in text-independent speaker verification." *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. Vol. 1. IEEE, 2001. |
| Original URL | https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.346&rep=rep1&type=pdf |
| Source type | Journal Article |
| Keywords | Gaussian Mixture Model - A model that expresses the probability density function of a random variable in terms of a weighted sum of its components (Fundamentals of SR)<br>RASTA - Abbr. for Relative Spectral Filtering. Filtering technique that increases its accuracy significantly in the presence of severe noise, but reduces the accuracy of the system in the absence of noise. (Fundamentals of SR) [don't use this filtering tech. in project] |
| Summary of key points (include methodology) | The problem that this paper aimed to address is that text-independent speaker verification (SV) systems perform worse when different languages are spoken. In the enrollment stage, a user provides voice data that the system analyzes and stores into a model (similar to how you input fingerprint multiples times for TouchID©). When a model is built off of voice data given in a certain language at enrollment, some vocal features change in the authentication stage so the system gives a False Reject decision. At the time of writing, this problem was hard to test because no databases contained enough data from speakers that spoke different languages. Despite this fact, the researchers chose to initially test the system with the 1998 NIST database, as well train the system with world language data from the 1996 language identification development database which provided 35 minutes of audio data for twelve languages. The SV system's design used Linear Predictive Coding (LPC) for feature extraction, RASTA filtering for normalization, and adaptive GMMs for modelling. In the testing phase, the researchers used two minutes of voice data, or ten utterances, to train the model. A control model was made that was created using American English speaker data. To train the model for different languages, two approaches were used. First was to train models for different languages independently and |

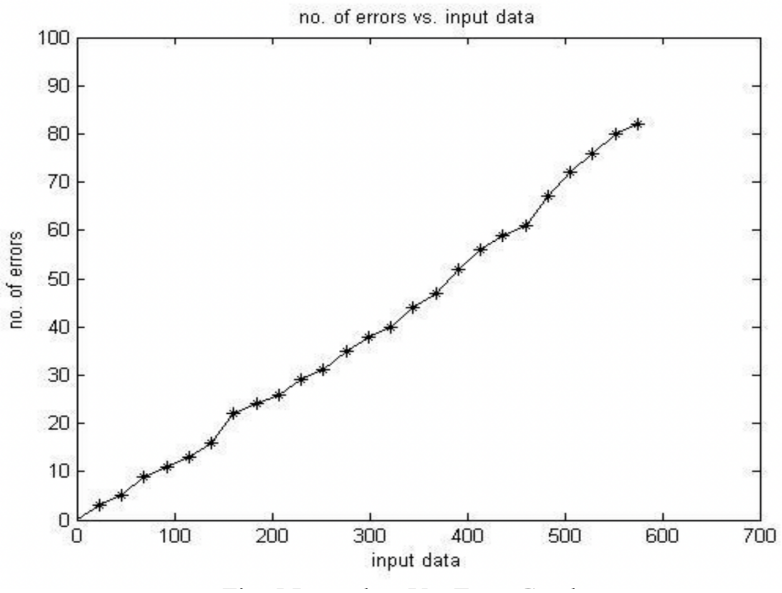
| | |
|---|---|
| | the second approach was to train a model with all different languages together using 64 and 256 mixtures. Using the first approach, the researchers found that the system performed considerably worse with east asian and southeast languages such as Vietnamese and Mandarin Chinese and slightly less worse with Romantic languages and Arabic and Tamil. Surprisingly, East asian languages like Japanese and Korean, along with German and Hindi performed similarly to the control model, and Farsi, a midde-eastern language, performed better than the control model. In the second approach, compared to the first, the pooled model, on average, performed slightly worse than the individual world language models and there was only a slight degradation in accuracy between using 64 and 256 mixtures. Through experimentation, the researchers came across some findings. The first is that the individual language model had a greater, but only slight increase in accuracy over a pooled model. The second was that with East Asian languages, the individual language model performed the worst with Mandarin and Chinese but performed as well as the control with Japanese and Korean. The third and final finding was that more data needed to be collected to improve the efficiency of the model which meant that at the time of publication, a better, and more expansive database was needed to improve language dependency in Text-Independent Speaker Verification (TISV). |
| Research Question/Problem/ Need | Problem: Speaker Verification systems based on Gaussian Mixture Models (GMMs) perform worse when different languages are spoken. |
| Important Figures | |
| Notes | ● This study was conducted in 2001 which means language may not be a problem in TISV with modern techniques.<br>● For results, the data was ranked but no listing of or in-depth explanation of statistical results was given or explained.<br>● I am not sure if this data is accurate because the authors put unknown explanations for unintended results.<br>○ For example, researchers wrote an unknown effect for Farsi model performing better than the control model. |
| Cited references to follow up on | NIST, The NIST 1996 Language Recognition Evaluation Evaluation, http://www.nist.gov/speech (Note: Visit more recent version) |
| Follow up Questions | 1) Would the system have performed better if MFCC feature extraction compared to LPC?<br>2) Since languages can be grouped into categories of origin and other similarities, would the system be able to perform proficiently if pool training was used for those groups as |

| | |
|---|---|
| | opposed to a variety of languages?<br>3) Do databases in 2021 have enough speaker data from different languages to improve world TISV systems as the researchers noted? |

# Article #8 Notes: Multilingual Speaker Recognition Using Neural Network

Article notes should be on separate sheets

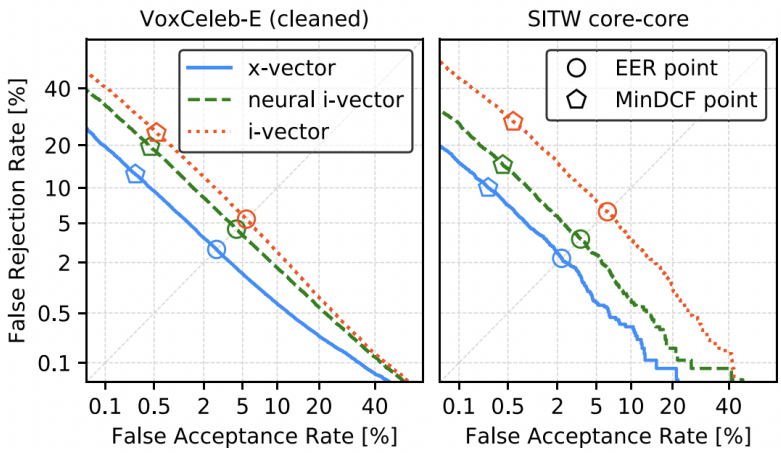
| | |
|---|---|
| Source Title | Multilingual Speaker Recognition Using Neural Network |
| Source citation (APA Format) | Kumar, R., Ranjan, R., Singh, S. K., Kala, R., Shukla, A., & Tiwari, R. (2009). Multilingual speaker recognition using neural network. *Proceedings of the Frontiers of Research on Speech and Music, FRSM*, 1-8 |
| Original URL | https://www.researchgate.net/profile/Rahul-Kala/publication/272086352_Multilingual_Speaker_Recognition_Using_Neural_Network/links/54da13200cf2970e4e7da5f2/Multilingual-Speaker-Recognition-Using-Neural-Network.pdf |
| Source type | Journal Article |
| Keywords | Back propagation Algorithm (BPA) - An algorithm that quickly calculates derivatives. Used in machine learning and ANNs to compute a gradient descent with respect to weights. (TechTarget)<br>Linear Prediction Coefficients (LPC) - Feature extraction method in audio signal processing.<br>Linear prediction Cepstral Coefficients (LPCC) - the coefficients of the Fourier transform illustration of the logarithmic magnitude spectrum. (InTechOpen)<br>Line Spectral Frequencies (LSF) - used to represent linear prediction coefficients (LPC) for transmission over a channel |
| Summary of key points (include methodology) | This paper, entitled, “Multilingual Speaker Recognition Using Neural Network,” was written by researchers at the Indian Institute of Information Technology and Management Gwalior, India. The engineering goal that this paper aimed to accomplish was to create a multilingual speaker verification system trained using neural networks. Before training the SV system, the researchers used a collection of 25 speaker utterances from a public domain in WAV format. In their design, they used utterances that had a format in which a vowel directly preceded a consonant in each word. The utterances were from both genders and were spoken in 5 languages: Hindi, Punjabi, Telugu, English and Sanskrit. Using a third-party software, CoolEdit, they preprocessed the signal to remove extraneous noise, and remove any silence from the audio data so only utterances remained. For feature extraction they used six types signal features: Linear Prediction Coefficients (LPC), Reflection Coefficients (RC), Linear Prediction Cepstral Coefficients (LPCC), Log area ratio (LAR), Arcus Sin Coefficients (ARCSIN) and Line |

| | |
|---|---|
| | Spectral Frequencies (LSF). Each one is calculated differently, but they provide a table of numbers that can each be formatted in a matrix. In the next step, the researchers used the data obtained from feature extraction and processed in a Back Propagation Neural Network to improve the efficiency of the model by fine-tuning the weights based on the error rate from the previous epoch. The Neural Network used 10,000 epochs meaning that it ran 10,000 times and it had 2 hidden layers. The results of the training were that out of 575 utterances, there were a staggering 82 errors meaning that the model only had 85.75% accuracy with a max performance of 91.3% and a min performance of 73.91%. The model accomplished the goal, however, the low accuracy rate shows that the model was ineffective at creating a *multilingual* SV system. |
| Research Question/Problem/ Need | Engineering Need: Creating a Text-Independent Speaker Verification that uses Neural Networks<br>Problem: Is a TISV system that uses Neural Networks effective for multilingual testing data? |
| Important Figures |  <br>**Figure 5** plots the data from Table 2 and shows the amount of input data vs. the number of errors. The graph is in a linear shape which means that the errors increased in a linear fashion. |

| | |
|---|---|
| Notes | ● Paper used LPCC only instead of MFCC which isn't the best feature extraction method.<br>● The paper aimed to create a multilingual SV system but they only tested it with languages of the same family.<br>○ There is no way to know if this type of model would be less or equally effective at giving a true accept/reject using other languages.<br>● A proper biometric test needs to differentiate errors between a False Accept and a False Reject rate. This paper just called it "errors," showing that there was a knowledge gap. |
| Cited references to follow up on | none |
| Follow up Questions | 1) Are SV systems created using statistical models (e.x. GMMs or HMMs) have a much higher accuracy (or higher TAR and FAR) compared to SV systems created using models built off of Neural Networks?<br>2) Would the model have an increased overall accuracy if the Neural Network trained each speaker independently as opposed to pool testing? |

# Article #9 Notes: ASVtorch Toolkit: SV with Deep Neural Networks

| | |
|---|---|
| Source Title | ASVtorch toolkit: Speaker verification with deep neural networks |
| Source citation (APA Format) | Lee, Kong Aik, Ville Vestman, and Tomi Kinnunen. "ASVtorch Toolkit: Speaker Verification with Deep Neural Networks." *SoftwareX* 14 (June 1, 2021): 100697. https://doi.org/10.1016/j.softx.2021.100697. |
| Original URL | https://link-springer-com.ezpv7-web-p-u01.wpi.edu/content/pdf/10.1007%2F978-0-387-77592-0.pdf |
| Source type | Journal Article |
| Keywords | Speaker Recognition, PyTorch, Deep learning |
| Summary of key points (include methodology) | Speaker Verification is a task that requires pattern classification and modeling. Although Speaker Verification can be done and has been tested with statistical models that are based on complex mathematics related to signal processing, this field greatly benefits from deep learning which can analyze any amount of audio data and can create a model with a greater accuracy. This paper was written by researchers from the Agency for Science, Technology and Research and also researchers from the University of Eastern Finland. The aim of the paper was to create a deep learning based Speaker Verification framework, which is just a platform for developing software applications, using PyTorch which is an open source machine learning library based on the Torch library. The Framework uses various methods for feature extraction including MFCC and i-vector and x-vector embedding for speaker embedding which is just "a fixed-dimensional representation of variable-length utterances" (Lee et Al, pg 2). If the framework is given an enrollment and test utterance, it can give a similarity score. The Framework contains different tools that each contain the Python packages needed to use them which streamlines the design process. The Framework contains two datasets included to develop ASV systems: VoxCeleb, which are more than a million utterances collected from youtubers, and Speakers in the Wild (STW). The researchers tested various systems using these datasets and created a Detection Error Tradeoff graph that showed that x-vector systems had the least tradeoff when it came to false rejection and false acceptance which is optimal compared to i-vector analysis which had a 2x higher |

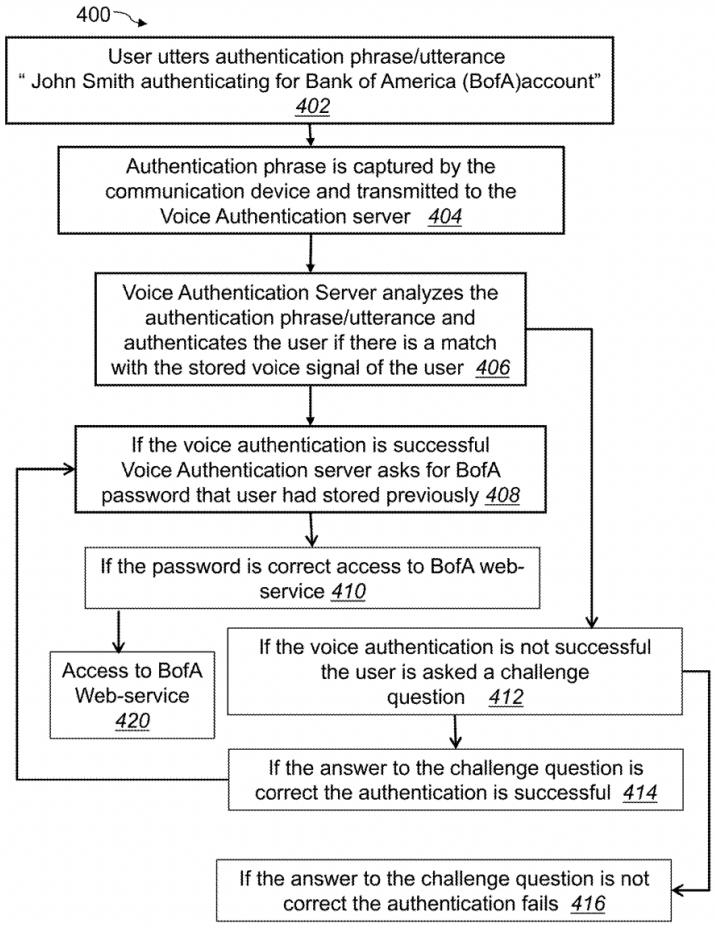
| | |
|---|---|
| | tradeoff. This toolkit accomplished its goal by creating a framework for creating Speaker Verification systems that is state-of-the art as it includes Neural Network tools, easy-to-use as it uses Python and Pytorch which are powerful languages/frameworks that are easier to code with it, and all-inclusive, as it contains all the packages needed to develop a Speaker Verification system built in. |
| Research Question/Problem/ Need | Engineering Need: PyTorch Framework for Speaker Verification |
| Important Figures |  <br>**Figure 5** shows a tradeoff curve between different systems tested with different datasets created using the ASVTorch toolkit. As can be seen, the x vector system has the lowest tradeoff curve while the i-vector system has the highest tradeoff curve. |
| Notes | • I don't know if I want to use Neural Networks for my TISV system but I think this toolkit will be useful if I plan on doing so.<br>• There doesn't seem to be many guides or a proper README so it would be hard to use. |
| Cited references to follow up on | 1) Zhu Y, Ko T, Snyder D, Mak B, Povey D. Self-Attentive speaker embeddings for text-independent speaker verification. In: Proc. Interspeech; 2018. p. 3573–7.<br>2) Zeinali H, Burget L, Rohdin J, Stafylakis T, Cernocky JH. How to improve your speaker embeddings extractor in generic toolkits. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2019, p. 6141–5, URL: https://github.com/hsn-zeinali/x-vector-kaldi-tf. |
| Follow up Questions | 1) How does this framework compare to other speaker verification toolkits in features and usability?<br>2) Do the researchers plan to continually update the toolkit to |

| | |
|---|---|
| | ensure that the libraries and packages used are up to date?<br>3) Does using a toolkit actually streamline the process of creating a Speaker Verification system or does it make it more time-consuming and arduous? |

# Patent #1 Notes: System and Method for Voice Authentication

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | System and Method for Voice Authentication |
| Source citation (APA Format) | Maddox, J., Smith, R. A., & Graylin, W. W. (2020). *(71) Applicant: ONVOCAL, INC., Northborough,MA*. 17. |
| Original URL | https://patentimages.storage.googleapis.com/d6/cd/e4/cd91d161aa5831/US10681032.pdf |
| Source type | Patent |
| Keywords | Single Sign-On (SSO - authentication method that enables users to securely authenticate with multiple applications and websites by using just one set of credentials. (onelogin) |
| Summary of key points (include methodology) | This patent entitled, “System and Method for Voice Authentication,” was written by Onvocal Inc. in 2020. This paper aimed to address a problem that users have to register their voice data for each and every account they enable voice authentication for. The goal of this patent was to create an architecture for a single sign on (SSO) voice authentication system that allows users to get access to their various third-party accounts using a central “voice print” model. The patent was created with the intent that companies could use this architecture to allow their users to securely navigate their services, on voice-input devices such as an Amazon Alexa or Google Home. In order for the SSO to work, a user must first create an account for the authentication server that contains registration data such as name, age, DOB, etc.. and also a list of internet accounts that the user would want to be used. The user provides voice data for the enrollment phase in two parts, general utterances and the numbers 0-9 which is used for a dynamic pin. And finally, the user can configure some general settings for the authentication server such as automatic log out time. In order to use the service, the user makes a request for a web service on a platform such as Alexa, Google Home, Siri, or etc. Then, the system determines if the request needs authentication, and it asks the user to recite a one-time pin. A model is built off the validation voice data which is compared by the system against the model created in the enrollment phase and it returns a result of Accept or Reject. The patent claims intellectual property to |

| | |
|---|---|
| | an exact design or implementation of the architecture previously mentioned with additional stipulations. The benefits of a product based on this patent are that it would make it easy for users to use VA as SSO, it can be developed on existing smart speakers, and users don't have to register their voice for each and every service they enable VA for. |
| Research Question/Problem/ Need | Engineering Need: Single-Sign On (SSO) for Voice Authentication |
| Important Figures |  <br>*FIG. 5*<br><br>**Figure 5** shows the elements of SSO for VA's architecture in a flowchart form. This is the final design listed in the patent. |

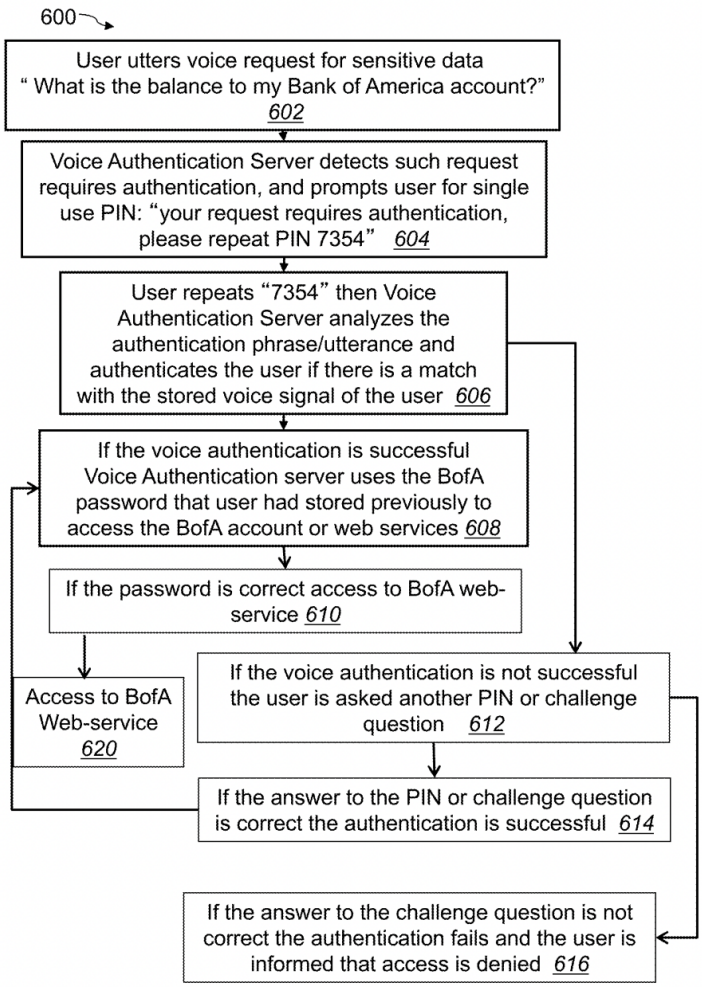



*FIG. 7*

**Figure 7** shows an example implementation of the architecture mentioned in the patent.

Notes

- This is a *very* general idea for a voice authentication system that doesn’t go into detail focuses mainly on:
  - Using voice authentication with AI-enabled devices like an Echo with Alexa or an iPhone with Siri.
  - Using cloud services to both store and authenticate a user.
- This seems to be aimed at VA for individual companies and their apps, services, etc…
  - The example they provided was a user saying an utterance that would allow them to be authenticated for a Bank of America account.
- Once users are authenticated, they can communicate voice commands to access various components of their accounts.

| | |
|---|---|
| | ○ The system ensures that the user provides the current pin/password.<br>● The goal of the Patent is to provide a sample architecture that lets users navigate various accounts for different companies and only using their voice. |
| Cited references to follow up on | n/a |
| Follow up Questions | 1) Wouldn't this system fail its general goal of improving user security by creating a backdoor for hackers to gain access to every one of a user's accounts?<br>2) Is a dynamic pin at the validation phase the most effective validation task if TISV was used in the enrollment phase? |

# Article #10 Notes: Voice Authentication Using Short Phrases: Examining Accuracy, Security and Privacy Issues

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Voice Authentication Using Short Phrases: Examining Accuracy, Security and Privacy Issues |
| Source citation (APA Format) | Johnson, R. C., Boult, T. E., & Scheirer, W. J. (2013, September). Voice authentication using short phrases: Examining accuracy, security and privacy issues. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (pp. 1-8). IEEE. |
| Original URL | https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6712713 |
| Source type | Journal Article |
| Keywords | Vaulted Voice Verification - System where a client device is validated if they can identify the difference between a regular and chaff model<br>Chaff Model - Model based on GMMs or HMMs that has been changed in a minor way |
| Summary of key points (include methodology) | This research paper was written by two researchers at the University of Colorado, R.C. Johnson and Terrance E. Boult, in 2013 at the IEEE Sixth International Conference on Biometrics. This paper researched how text-dependent and text-independent models can be used to improve the accuracy and security of SV systems. They used a new protocol at the time called Vaulted Voice Verification (VVV) to accomplish this. Vaulted Voice Verification has a different way of authentication both in enrollment and validation. The goal of this project was to test the researcher's new model compared to traditional GMMs, and show that it was more accurate in validating a speaker. In enrollment, the system uses statistical models to create a model from speaker data which is similar to most systems, but in addition, it creates chaff models. Chaff models are models created based on regular models (this is how I will refer to them) but they have one minor difference. It's important to note that there isn't a standard to set chaff models and the authors didn't mention how it was changed. Anyways, in the validation phase, the system sends both the regular model and the chaff model to the client device, and it is the client's goal to find the difference between them. The reason |

| | |
|---|---|
| | chaff models are created is because the client device can find the difference between both models using the voice data of the legitimate speaker. However chaff models are nearly identical to the regular model, so an attacker would not be able to find the difference. In their design, for the extensive revision made to VVV, they decided to use a new type of challenge. In the models they tested, the researchers used challenges where the speakers said predictable responses i.e yes, no. However, if a response is predictable, this compromises the security of the system by allowing an attacker to know what data to get from a speaker. So, they needed a challenge that could give predictable responses, to some degree, to ensure the speakers were giving clear data, and so the validation phase wouldn't take too long. They came up with a challenge where the system gives a picture, and the user has to describe it in one-word responses. The system knows that the responses will be short, but the attacker cannot predict what the speaker is going to say. When talking about this VVV system, the researchers stressed how security was penultimate and gave a sample architecture that ensured that data sent between a server and client device was encrypted. For training, they used a speech corpus or database of 48 speakers which contained audio recordings of short phrases in 20 minute sessions. In all designs, the researchers used Gaussian Mixture Models (GMM) to create models from the voice data. In experimentation, they tested four different designs: traditional text-dependent GMM model where a user just repeats phrases, VVV with binary questions that have two responses, VVV with multiple choice answers, and VVV with text-independent prompts. In this experiment, they defined text-dependent to refer to a system relying on predictable responses and text-independent to refer to a system not being able to predict the output, and solely relying on feature extraction to validate a user. From the results of the experiment they found control had an Equal Error Rate (EER) of 8%, the binary response VVV had an EER of 6%, the m.c. VVV had an EER of 4%, and the extensions made to the VVV had an EER of just 1%. |
| Research Question/Problem/ Need | • Engineering Need: Making Modifications to a Vaulted Voice Verification (VVV) System to improve security. |

Important Figures

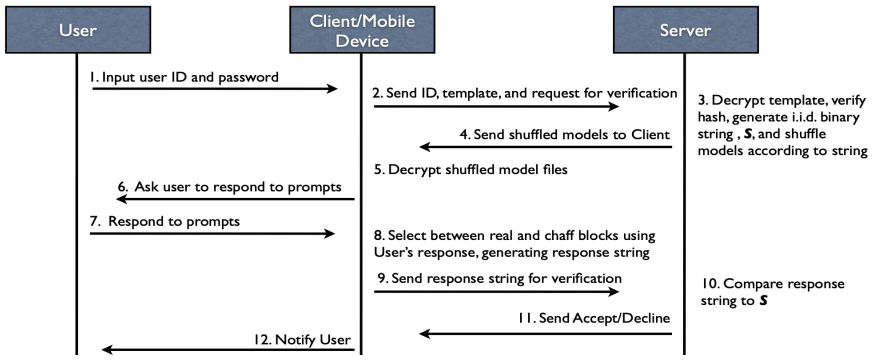



**Figure 3** shows the architecture behind the validation phase in a VVV system. The encryption in each step ensures that there is a secure connection between the client and the server.


**Figure 4** shows an example of how a user may respond to the novel challenge prompt created in this experiment.

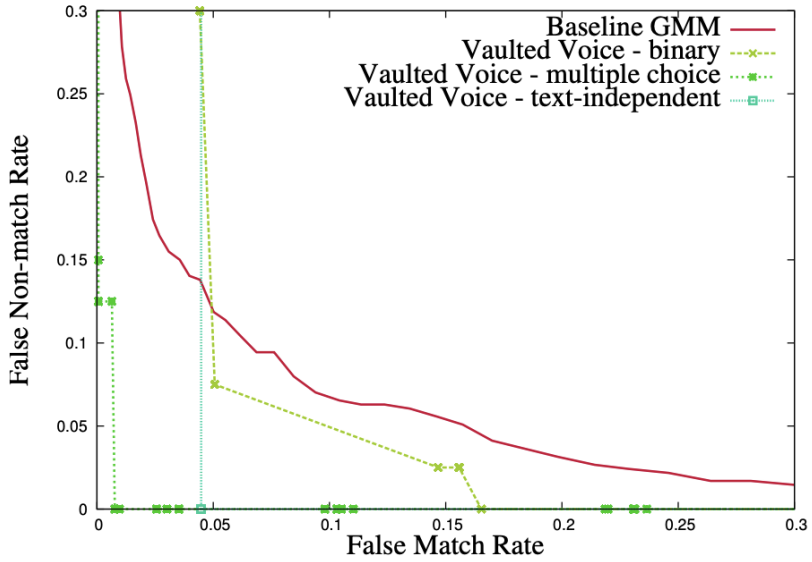



**Figure 5** shows a Linear Scale Data Error Trade Off Plot which compares the False Acceptance Rate in the x-axis to the False

| | |
|---|---|
| | Rejection Rate in the y-axis. The measures are of percentages. This has to be shown in a linear plot because scores of 0 cannot be shown in a logarithmic plot. |
| Notes | ● Engineering Need: Making Modifications to a Vaulted Voice Verification (VVV) System to improve security.<br>● The goal of this paper was to<br>○ increase security of VVV by incorporating Text-Independent Speaker Models<br>○ Increase security of system per question<br>○ Provide a security analysis of VVV protocol using attack models.<br>○ Analyze the accuracy and security trade-off from TISV models.<br>● Vaulted Voice Verification (VVV) - a text dependent challenge-response-based verification protocol. **Generates two types of models**: A normal one and a chaff model.<br>○ Normal Model:<br>■ Uses Guassian Mixture Models (GMMs)<br>● Which is a function that represents subpopulations of data within an overall population. Hence it's a "mixture model."<br>■ Hidden Markov Models (HMMs)<br>● A statistical model used when one process depends on another.<br>○ Chaff Model:<br>■ Created by altering the Normal model.<br>■ There isn't a standard for this and the authors described deciding how to alter the data as "more art than science."<br>○ Why is this used?<br>■ This idea is that only a model created from a validated user at enrollment, can be used to differentiate between the real and chaff model.<br>■ An attacker wouldn't be able to supply the proper voice data to do this.<br>● Enrollment<br>○ RSA Encryption is used to ensure there is a safe connection between client and server.<br>○ Build a model for **each** phrase.<br>○ In the original protocol: System asks the user to repeat a series of phrases.<br>○ In the extended protocol: the System asks Text-Dependent and Text-Independent prompts.<br>○ In order to increase security, users are supplied an image, and they are tasked with describing. |

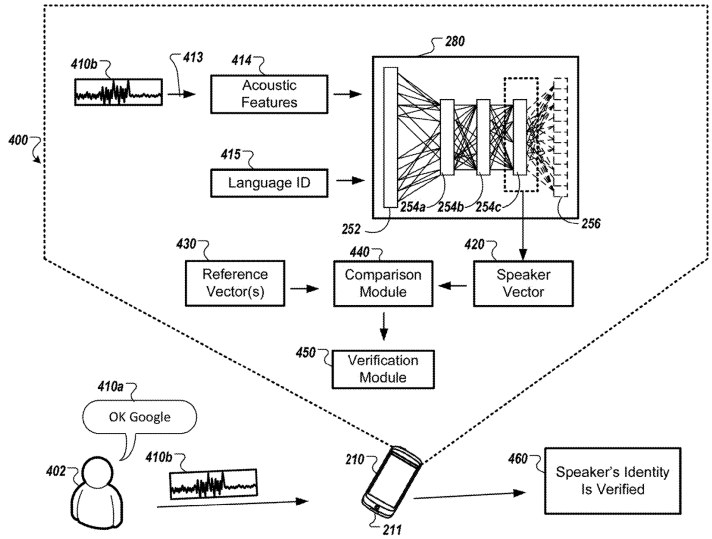
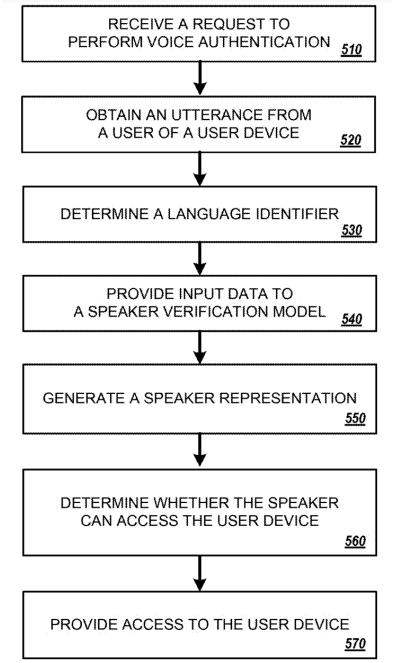
| | |
|---|---|
| | ■ This allows users to give short responses.<br>■ It ensures that different users won't say the same phrase.<br>● Text Dependent and Text Independent<br>○ *What's the purpose?*<br>■ Text-dependent models provide a lot of data more easily<br>■ Text-independent models prevent replay attacks where an attacker just plays an audio recording.<br>● The authors consider their implementation 3 factor authentication because it tests something you have, know and are. You need an authorized phone to connect with the system, you have an ID number that the system checks the user against, and the biometric used is your voice.<br>Experiment:<br>● Used a database of audio files (speech corpus) composed of 48 speakers with 22 male and 28 female speakers.<br>● Each speaker had an imposter, or a different speaker, selected for training.<br>● They used "all in all" or pooled testing where every speaker is used in training the model.<br>● They had a control using GMMs where the user just repeated a series of phrases meaning there wasn't any privacy or security in this type of model.<br>● In the second round, the tested VVV with binary responses where there were only two answers for each question asked.<br>● In the third round, they tested VVV by asking multiple choice questions.<br>● In the fourth and final round, they tested using the enhanced model. |
| Cited references to follow up on | 1) F. Alegre, R. Vipperla, N. Evans, and B. Fauve. On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. 20th European Signal Processing Conference (EUSIPCO 2012), 2012.<br>2) M. Just and D. Aspinall. Personal choice and challenge questions: A security and usability assessment. Proceedings of the 5th Symposium on Usable Privacy and Security, 8, 2009.<br>3) T. Kinnunen, Z. Wu, K. Lee, F. Sedlak, E. Chng, and H. Li. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages 4401–4404. IEEE, 2012. |

| | |
|---|---|
| Follow up Questions | 1) Why didn't the researchers test the VVV with prompts which aren't confined to one-word responses?<br>2) Is VVV more accurate compared to Authentication Methods based on HMM or Time Delay Neural Networks?<br>3) If a hacker constructs a voice reconstruction model based on a registered user's voice data, wouldn't they be able to easily hack into the system since it requires short input data? |

# Patent #2 Notes: Speaker verification across locations, languages, and/or dialects

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Speaker verification across locations, languages, and/or dialects |
| Source citation (APA Format) | Moreno, I. L., Wan, L., & Wang, Q. (2021). *Speaker verification across locations, languages, and/or dialects* (United States Patent No. US11017784B2). |
| Original URL | https://patentimages.storage.googleapis.com/52/b9/ea/c1cbd6c698f091/US11017784.pdf |
| Source type | Patent |
| Keywords | Vector - A 1-dimensional array of numbers used in ML to conveniently organize data<br>Utterance - An uninterrupted chain of spoken language.<br>Language-Independent Speaker Verification |
| Summary of key points (include methodology) | This patent was written by Ignacio Lopez Moreno, Li Wan, Quan Wang and its assignee was Google. This patent claimed intellectual property to a unique Speaker verification (SV) system they referred to as “language-independent” SV which uses neural networks. The architecture of this SV system has been seen before, however, for the unique languages, it uses a “hotword” or a language identifier which is a word that has been previously set by the user that allows for the neural network to recognize the language. This is done so that if the same speaker speaks to the system in different languages, it will use a different language model to validate the user. In a traditional system, additional training for speakers who speak different languages is a face that hasn’t been recognized. |
| Research Question/Problem/ Need | Engineering Need: Text Independent Speaker Verification (TISV) system for different languages and dialects. |

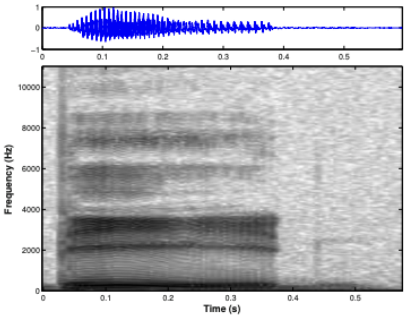
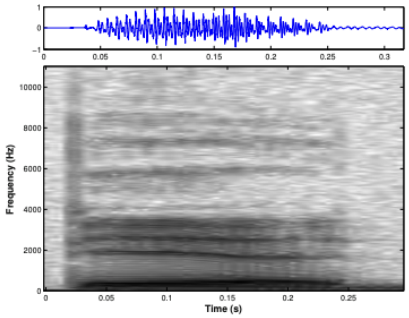
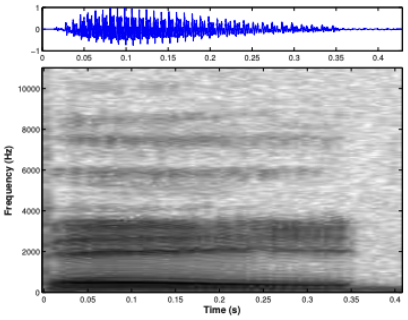
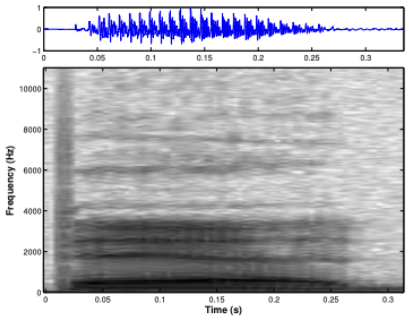
| | |
|---|---|
| Important Figures |  <br>FIG. 4<br>**Figure 4** shows a sample system based on the patent's claims. Data is sent through a mobile device where extracted features and language ID are sent to a neural network and a speaker vector is returned. The verification module compares the reference and speaker vector to return a boolean result of accept or reject to validate the user's identity.<br>  <br>FIG. 5<br>**Figure 5** is a flowchart that shows the validation process of a *language-independent* speaker verification system. The unique step is the third one which is the system determines a language identifier in the form of a "hotword." |
| Notes | • This paper, like other patents I have read on SV systems, |

| | |
|---|---|
| | provides a *very* general architecture for an SV system but it focuses on different languages which is the same focus as my project.<br>● They call it a “language-independent” model to convey that it provides accurate validation regardless of the language.<br><br>General Patent Claims for an SV system that:<br>● receives an utterance of a predetermined hotword for a language or location associated with the user.<br>● uses a neural network for training and is trained with different languages or dialects.<br>● provides personalized responses to the user based on the user's identity. |
| Cited references to follow up on | n/a |
| Follow up Questions | 1) Would simply recognizing the speaker’s language improve the accuracy of an SV system?<br>2) Is recognizing a hotword the most practical way to recognize a speaker’s language/dialect if a proposed system uses short utterances?<br>3) Can a hotword be used to recognize a speaker’s *dialect* if a speaker doesn’t use dialect specific phrases or slang? |

# Article #11 Notes: Phonetics and Phonology

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Phonetics and Phonology (Chapter 4) |
| Source citation (APA Format) | Beigi, H. (2011). *Fundamentals of Speaker Recognition*. Springer US. https://doi.org/10.1007/978-0-387-77592-0 |
| Original URL | https://link-springer-com.ezpv7-web-p-u01.wpi.edu/content/pdf/10.1007%2F978-0-387-77592-0.pdf (Pg. 167) |
| Source type | Book Chapter |
| Keywords | Phonetics, Phonology, Phonemes, Phones, Allephones (defined in notes) |
| Summary of key points (include methodology) | "Phonetics and Phonology" is the fourth chapter in the book entitled, *Fundamentals of Speaker Recognition*, which was written by Homayoon Beigi. Phonetics is the study of sounds produced by the human vocal system and those individual parts can be divided into phonemes and phones. Phonemes are semantically significant sounds that occur in a language while phones are elementary sounds that are the smallest segment of speech. As an example, the sound p is considered a phoneme in English, but the same phoneme has two phones, *p* and $p^H$. There are two phones because you pronounce $p^H$ with a puff of air, and pronounce *p* without that puff. This is important in text dependent applications of SV systems because although some letters in the validation text are the same from the training data, if they are different phones, they will sound different from the computer's perspective. Phonology is the study of phonetics at a language level. From a signal processing perspective, grammar and sentence structure aren't very important, but the phones and phonemes in different languages are. For some context, although English has 26 letters, it has 44 phonemes. 70% of languages have between 20 and 37 phonemes. This is important in speaker recognition because languages with more phonemes need more enrollment data to ensure that prompts spoken in the validation phase were previously analyzed by the system. Vowels are a type of phoneme that are easy to recognize and they also can be seen by a computer in formants on a spectrogram. They are important in speaker recognition as having enrollment and validation data with a significant amount of vowel utterances can increase the accuracy of the system as there will be less of them compared to consonants throughout most languages. Many other factors can impact the |

| | |
|---|---|
| | frequency response of spoken language, such as pitch, tone, and whispering, which are factors that should be accounted for in the design of a SV or SR system. |
| Research Question/Problem/ Need | RQ: “How are languages different from a signal processing perspective?” |
| Important Figures |  <br>**Fig. 4.10:** bead /biːd/<br>(In an American Dialect of English)<br>  <br>**Fig. 4.11:** bid /bɪd/<br>(In an American Dialect of English)<br>  <br>**Fig. 4.12:** bayed /beɪd/<br>(In an American Dialect of English)<br>  <br>**Fig. 4.13:** bed /bɛd/<br>(In an American Dialect of English)<br><br>Figures, **4.10, 4.11, 4.12, & 4.13,** show spectrograms for 4 different words that sound phonetically similar. Although the words, *bead*, *bid*, *bayed*, and *bed*, sound similar, they produce different waveforms, or signal shapes, and also different formants, or peaks in frequency. |
| Notes | • Phonetics - The study of sounds produced by the human vocal system.<br>• Phone - Elementary Sounds that occur in languages. They are the smallest segment of speech with distinct vocal patterns.<br>• Phoneme - Semantically significant sounds that occur in a language.<br>• Allophone - Different phones which convey the same phonemic information.<br>  ○ In English, p can be pronounced in 2 different ways. P |

aspirated with air is written as pH while a regular p is just written as p. Although p and pH can be used interchangeably, in other languages, that sound difference can change the meaning of the word itself. **So p and pH are different phones.**

- The four main elements of speech production are: initiation, phonation, articulation, and coordination.

1) **Initiation**:
   - Initiation - A function of the airstream mechanism and the direction of airflow.
   - Can be either pulmonic, glottalic, or velaric.
     - Pulmonic - Initiated by lungs
     - Glottalic - Initiated by Larynx
     - Velaric - Initiated by tongue.
   - Air can move inwards, *ingressive,* or it can move outwards*, egressive*.
2) **Phontation**:
   - Phonation - Deals with acoustic energy generated by vocal folds at the larynx.
   - Can be *unvoiced*, *voiced*, and *whisper*.
3) **Articulation**:
   - Deals with the place of articulation, degree of structure, and the aspect of articulation.
   - These result in distinct types of sounds such as stops, trills, and resonants which make up different phonemes.
4) **Coordination**
   - Deals with the way you move your organs to produce an *advanced* sound. Related to initiation, phonation, and articulation.

- Vowels
  - The tongue is the most important body part used in making a vowel.
  - There can be voiced and voiceless vowels.
  - **IMP**: Important in speaker recognition because:
    1) They are easy to recognize since they are voiced differently.
    2) Can be seen in formants.

**Phonology and Linguistics:**

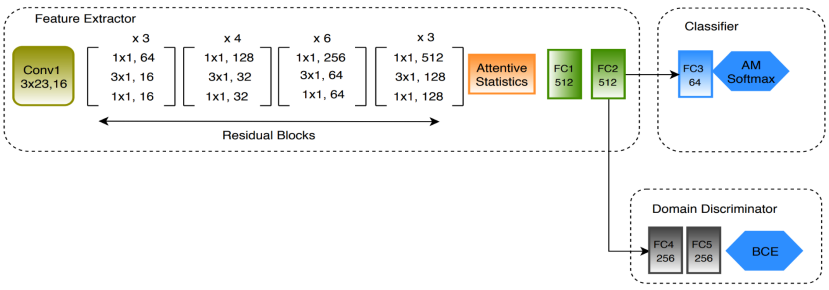
- Phonology - study of phonetics in the framework of specific languages.
- **Phonemic Utilization Across Languages**
  - A study conducted in 1984 across 315 languages found that 70% contained between 20 to 37 different phonemes.
  - For some context, some Hawaiian languages contain

| | |
|---|---|
| | only 11 phonemes (6 con. and 5 vow.) while some !Kung languages famous for their clicking noises contain 141 phonemes.<br>○ There are 39 phonemes in English. |
| Cited references to follow up on | None. I won't be going in-depth into phonetics with my project. |
| Follow up Questions | 1) How will increasing enrollment data impact the accuracy of an SV system across different languages if they have different phonemes?<br>2) Will an SV system that uses prompts that contain a significant amount of vowels have a greater accuracy compared to prompts with many vowels? |

# Article #12 Notes: Adapting End-To-End Neural Speaker Verification To New Languages And Recording Conditions With Adversarial Training

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Adapting End-To-End Neural Speaker Verification To New Languages And Recording Conditions With Adversarial Training |
| Source citation (APA Format) | Bhattacharya, G., Alam, J., & Kenny, P. (2019). Adapting End-to-end Neural Speaker Verification to New Languages and Recording Conditions with Adversarial Training. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6041–6045. https://doi.org/10.1109/ICASSP.2019.8682611 |
| Original URL | https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8682611 |
| Source type | Conference Paper |
| Keywords | 1) Speaker Verification<br>2) Adversarial Training - A collection of techniques to train neural networks on how to spot intentionally misleading data or behaviors. Goal is not just to identify but to predict vulnerabilities. (DeepAI)<br>3) Domain Adaptation - Aim is to train a neural network on a source dataset and secure a good accuracy on the target dataset which is significantly different from the source dataset. (towardsdatascience)<br>4) End-to-End (E2E) - refers to training a possibly complex learning system represented by a single model (specifically a Deep Neural Network) that represents the complete target system, bypassing the intermediate layers usually present in traditional pipeline designs. (towardsdatascience) |
| Summary of key points (include methodology) | This conference paper was written for the 2019 IEEE Conference on Acoustics, Speech and Signal Processing in 2019 by researchers at McGill university. The research problem that this paper aimed to address was, Can a speaker recognition system, modelled using |

| | |
|---|---|
| | neural networks be trained using end-to-end learning? Their goals were to create a novel architecture for extracting speaker embeddings in a system called Domain Adversarial Neural Speaker Embeddings (DANSE), show that DANSE could be optimized using E2E, and add adversarial training into the model to learn domain invariant features. The design of the DANSE model consisted of 3 components: the feature extractor, the classifier, and the discriminator. The feature extractor looked at a collection of 10, five second long samples of audio from each recording and extracted 23 MFCC features from each frame. The features vectors were sent *simultaneously* to the classifier and the discriminator. The classifier's goal was to find the similarity between the two audio samples and the discriminator's goal was to find the difference. If both models returned the same result, then the final system returned a result of acceptance or rejection. Current state of the art models include i-vector and x-vector based approaches which use PLDA classifiers which take more time and are more resource heavy. The results showed that even compared to those models, the DANSE model with just cosine similarity as classifier, had the lowest pooled data Equal Error Rate (EER) at 13.29% in the NIST SRE. The DANSE Model also had the lowest EER in the Speakers in the Wild (STW) dataset at 8.32%. The findings of this paper were that adversarial learning was perfect for this task as the main point of speaker verification is to find if two audio samples are *different* from each other. The second finding was that the main challenge of this architecture was finding the optimal similarity which was taken care of by the DNN. The third finding was that the DANSE model with E2E and a simplistic Cosine classifier, was more time and power efficient compared to other NN SV architectures. |
| Research Question/Problem/ Need | Problem: Can a Speaker Recognition System, modelled using Neural Networks (NN), be trained using End-to-End learning? |
| Important Figures |  <br>**Figure 1** shows a flowchart of the DANSE model. |

| *Model* | *Classifier* | *Cantonese* | *Tagalog* | *Pooled* |
|---|---|---|---|---|
| i-vector | PLDA | 9.51 | 17.61 | 13.65 |
| x-vector | COSINE | 36.44 | 41.07 | 38.69 |
| **x-vector** | **LDA/PLDA** | **7.52** | **15.96** | **11.73** |
| x-vector | PLDA | 7.99 | 18.46 | 13.32 |
| AMS | COSINE | 11.44 | 21.22 | 16.28 |
| DANSE | COSINE | 8.84 | 17.87 | 13.29 |

**Table 1** shows the performance of different SV system architectures when trained with the NIST-SRE 2016 dataset.

| *Model* | *i-vector* | *x-vector* | *AMS* | *DANSE* |
|---|---|---|---|---|
| ***EER*** | 11.47 | 10.51 | 9.87 | **8.32** |

**Table 2** shows the performance of different SV system architectures when trained with the STW dataset.
**IMP**: No graphs were provided.

Notes

**Introduction:**

- Text-Independent Speaker Verification - binary classifiers that given two recordings answer the question: Are the people speaking in the two recordings the same person?
- Verification Score - Answer of an SV system
  - (Sorted by Strength)
  - Distance Metrics
    - Mean Squared Error
    - Cosine Distance
  - Likelihood Ratio (This is the one I'm using)
  - i-vector/PLDA
  - I-vector with Deep Neural Network (DNN)
- Goals
  1) Create a novel architecture for extracting speaker embeddings called Domain Adversarial Neural Speaker Embeddings (DANSE).
  2) Show that the DANSE model can be optimized E2E for similar performance to cosine scoring.
  3) Add Adversarial Training into the Speaker Embedding model to learn domain invariant features.
- Modern Datasets:
  - NIST-SRE 2016 and Speakers in the Wild (STW)
  - Difficult to work with because target data isn't available for verification systems specifically.

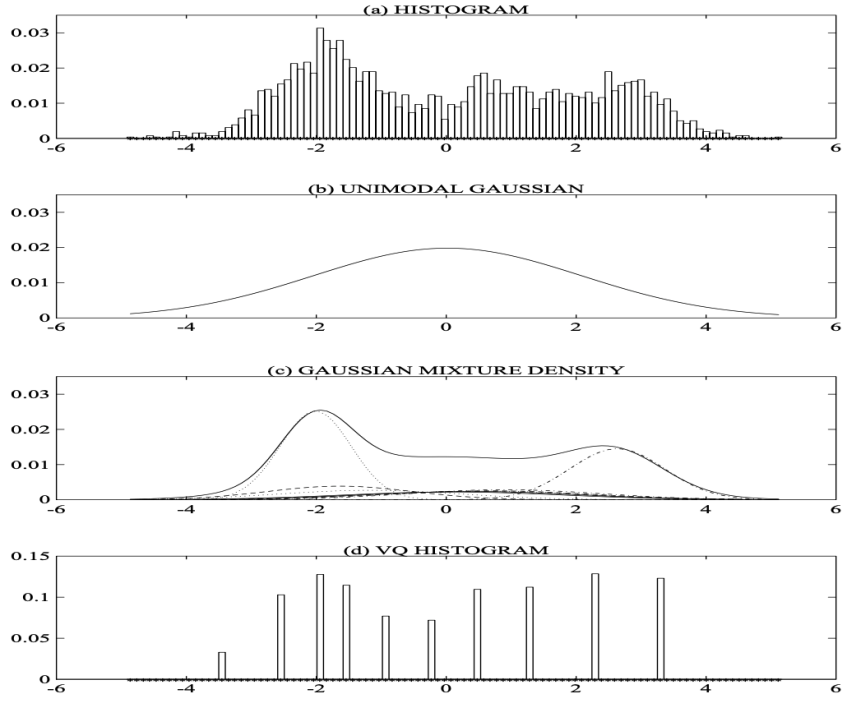
| | |
|---|---|
| | ○ This causes degradation in performance<br>○ So, the researchers added their third goal to navigate around that roadblock and speed up performance.<br>● Main Objective:<br>○ Speaker Recognizers can be learned E2E or in a NN model that does everything.<br>○ This means that the researchers don't have to code for each and every component such as preprocessing, feature extraction, etc. All they have to do is code for that within the NN model which is more complex, but results in a simplified development pipeline.<br>● Gained Knowledge:<br>○ Main challenge with this is optimization of similarity.<br>■ **IMP**: I need to find a way to do this as it is impractical to calculate threshold value every time.<br>○ This type of task is perfect for adversarial training as the whole point of speaker verification is to analyze audio data to see if it belongs to a registered user.<br><br>**Experimentation**:<br>● Training Data:<br>○ NIST-SRE evaluations (2004-2010)<br>○ Switchboard Cellular audio<br>● Feature Extraction:<br>○ 23-dimensional MFCC features (As of Oct. 21 I am using 40 MFCC features)<br>● Feature Extractor Model:<br>○ 48 Input Layers<br>○ Attentive statistics layer & 2 Fully Connected Layers<br>● Classifier Model<br>○ 1 Hidden Layer<br>○ AM-Softmax Output Layer<br>● Discriminator Model<br>○ 2 Hidden Layers<br>○ Binary Cross-Entropy (Bce) Output Layer<br>● Optimization:<br>○ Optimized Each Model Separately for Max. Performance<br>● Sampling:<br>1) Split random chunks of audio around 5 seconds from each recording.<br>2) In an epoch, each recording is sampled 10 times.<br>● Speaker Verification:<br>○ Scoring Method: Cosine Distance<br>● Performance:<br>○ Equal Error Rates (EER) |

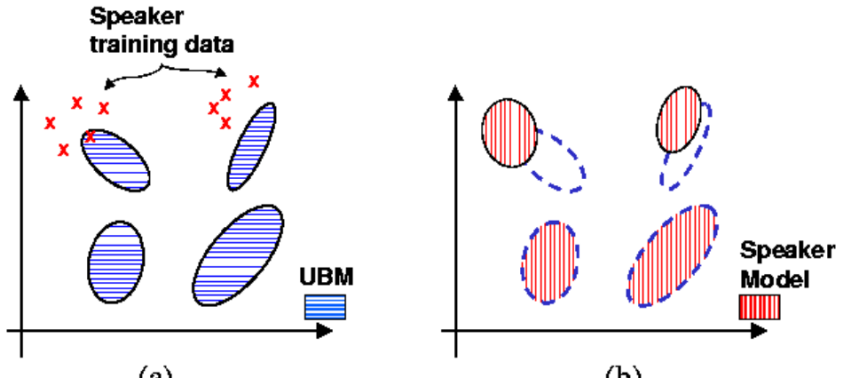
| | |
|---|---|
| Cited references to follow up on | 1) Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, “Generalized end-to-end loss for speaker verification,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4879–4883.<br>2) Gautam Bhattacharya, Jahangir Alam, and Patrick Kenny, “Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification,” in Acoustics Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on. IEEE, 2019. |
| Follow up Questions | 1) Why weren’t any graphs such as Data Error Tradeoff (DET) Plots provided if the system was so accurate?<br>2) Since the system uses cosine similarity for classification which is fairly simplistic, would the DANSE model perform better, albeit with more computing time, using a more advanced classifier such as Probabilistic Linear Discriminant Analysis? |

# Article #13 Notes: Gaussian Mixture Models

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Gaussian Mixture Models |
| Source citation (APA Format) | Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*, 659-663. |
| Original URL | link |
| Source type | Review Article |
| Keywords | GMM, Mixture model, Gaussian mixture density |
| Summary of key points (include methodology) | A Gaussian Mixture Model (GMM) is a probability density function (PDF) that is simply a weighted sum of Gaussian distributions, also known as bell curves. This is the equation for a GMM $p(x\|\lambda) = \sum_{i=1}^{M} w_i g(x\|\mu_i, \sum_i)$. $x$represents continuous data in vector form. , $w_i$represents the weights used to “fine-tune” the model in training, and the sigma notation is used to add the number of Gaussian densities from 1 to the number of mixtures, $M$. GMMs are most commonly used in biometric systems, especially speaker verification systems, or any type of task that provides a continuous feature set. Audio is a continuous function of time vs. amplitude, but computers use sampling to section off that function, and only store selected values to represent the sound. The numbers are stored in a two dimensional vector or list which can be represented as an electrical signal which can move a diaphragm on an electronic speaker. The reason why GMMs are used in biometric systems is because they are able to represent many subpopulations of data in one large distribution. A Vector Quantizer (VQ) is a comparable modelling technique that “compresses" a histogram, or a large vector, into a certain number of values. VQ can model data, but it doesn’t provide a distribution, it just provides a smaller vector, and it also doesn’t do well at representing subpopulations. A unimodal gaussian distribution is essentially using a GMM with just one mixture. This method only works best with datasets that only have one subpopulation because they look at the probability density of the dataset as a whole. Gaussian Mixture Models must be trained which is the process of adjusting weights to get a model that does a better job of representing the original dataset. There are two methods for training: Expectation-Maximization (EM) or Maximum A Posteriori (MAP) estimation. EM is an iterative, and perhaps brute force way to find the maximum likelihood estimates for model parameters in |

| | |
|---|---|
| | model training. The process of an EM algorithm is to estimate values for each mixture based on the current model to create a new model. If that new model maps the dataset better, it becomes the new model and the process repeats again. Maximum A Posteriori (MAP) Parameter estimation is similar to EM, as it is an iterative algorithm that first estimates values. However, before the second step, if values for a model are estimated which offer a better accuracy, the algorithm, rather than use that model, creates another mixture of the data from the old model, and also the new estimates. This is done as a failsafe, of sorts, for when the algorithm cannot estimate values to create a better model. In this case, the algorithm will rely less on the new estimates, but will rely more, but not entirely, on the old, "sufficient" data. |
| Research Question/Problem/ Need | RQ: What is a gaussian mixture model? |
| Important Figures |  **Figure 1** shows a comparison of 4 types of distribution modelling for a feature vector of cepstral coefficients extracted from a male utterance of 25 seconds. |

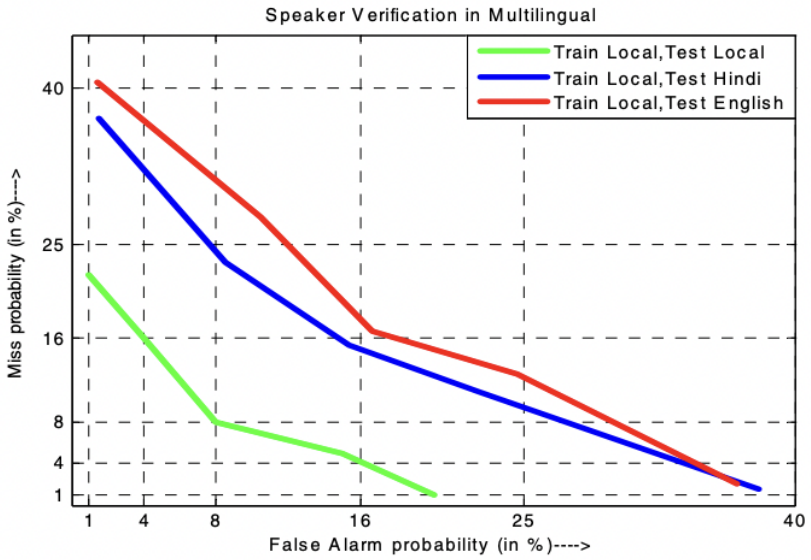
| | |
|---|---|
| |  <br>**Figure 2** is a graphical representation of fitting/training a GMM-UBM model. |
| Notes | • Gaussian Mixture Modelling is not the best method for Speaker Verification systems anymore, but it is less computing intensive, and can work well with proper implementations.<br>• I used Expectation Maximization (EM) in my Proof of Concept #1 so I will look into changing the training method to use MAP.<br>• I just used a apples-to-apples comparison in my Proof of Concept #1 but a better way is to develop a Universal Background Model. |
| Cited references to follow up on | Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models. IEEE Transactions on Acoustics, Speech, and Signal Processing 3(1) (1995) 72–83 |
| Follow up Questions | 1) How does adding more mixtures to a GMM affect performance?<br>2) What is the lowest, sensible, amount of enrollment data that can be used in an SV system using GMM? (Background for this question is that GMM based SV systems have been known to require a lot of data–don't know that number is. ) |

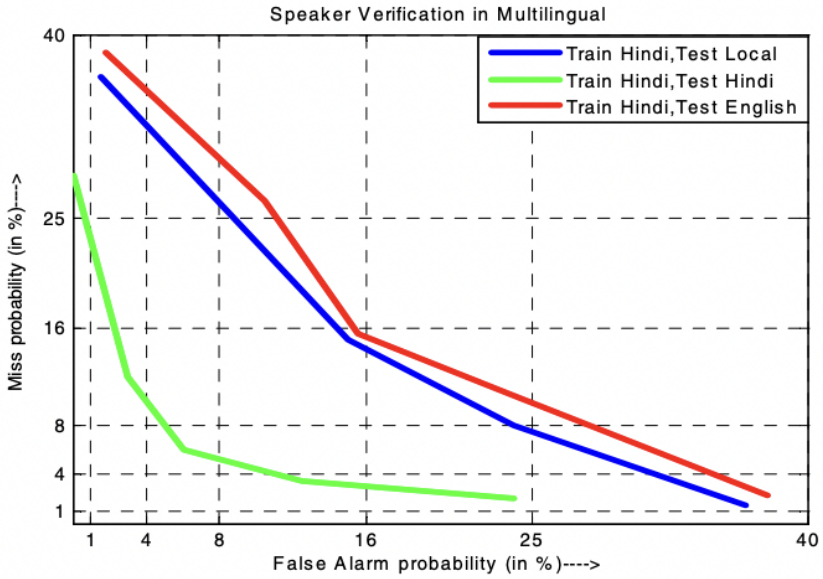
# Article #14 Notes: A Multilingual Speech Database for Speaker Recognition

Article notes should be on separate sheets

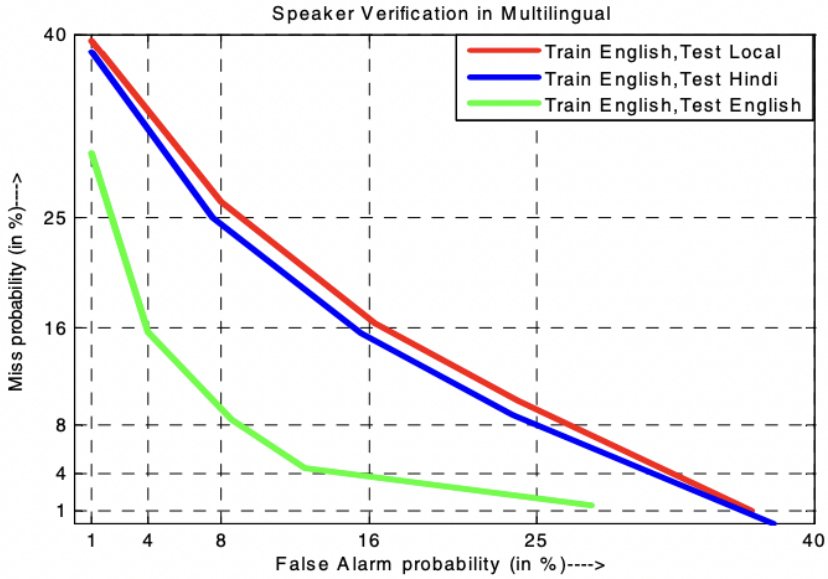
| | |
|---|---|
| Source Title | A Multilingual Speech Database for Speaker Recognition |
| Source citation (APA Format) | Bhattacharjee, U., & Sarmah, K. (2012). A multilingual speech database for speaker recognition. *2012 IEEE International Conference on Signal Processing, Computing and Control*, 1–5. https://doi.org/10.1109/ISPCC.2012.6224374 |
| Original URL | https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6224374 |
| Source type | Conference Paper |
| Keywords | Speaker Recognition, Multilingual Speech Database, Gaussian Mixture Modeling |
| Summary of key points + notes (include methodology) | This paper entitled, “A Multilingual Speech Database for Speaker Recognition,” was written by Computer Science and Engineering researchers at the Rajiv Gandhi University in Arunchal Pradesh, India in 2012. The problem that this paper addressed is that a user speaking in different languages impacts the authentication accuracy of Speaker Verification systems. This is a problem in countries like India where a majority of people are bilingual or multilingual and would prefer to interact with technology in different languages. The first goal of this paper was to create a speech corpus, or audio database, of recordings from speakers who spoke three different languages, English, Hindia, and Arunchali local languages. The second goal of this paper was to create a GMM-UBM based speaker verification system and train it to achieve highest accuracy with different languages in the database, but also with different microphones. The first step the researchers took was collecting recordings for the audio database. Languages from Arunchal Pradesh are called Tani languages and they fall under the Tibeto-Burman language family. This differs from the other languages they tested, Hindi and English, which both fall under the Indo-European family. The researchers chose four specific Arunachali languages: Adi, Nyishi, Galo and Apatani, based on popularity. For the DB, the researchers recorded four to five minutes of data from speakers in the 20-50 age category. The speakers read |

| | |
|---|---|
| | a grade school story and participants used four different types of microphones ranging from 16kHz to 44.1kHz in the sampling rate. In developing the SV system, the researchers chose a Gaussian Mixture Model - Universal Background Model (GMM-UBM). The framed the audio in 6.315s intervals (rough approximation I calculated) and extracted 19 MFCCs per frame to create a 38X19 feature vector. The Universal Background Model, which is a mixture of different feature vectors, contained 1024 mixtures. In experimentation, the researchers set the enrollment data length to two minutes, with 15, 30, and 45, second samples all tested in the validation phase. The results generally support the hypothesis that speaking different languages from the given enrollment decreases authentication accuracy but it also shows this isn't the case for all languages. The researchers found that when a system was trained with a local language, the system decreased in authentication accuracy, on average, by 7.95% when the user spoke in English or Hindia. Similarly, when the system was trained with Hindi, the authentication accuracy decreased by 9.25% when the user spoke a native language or English. However, when the user spoke in English for training, the researchers found that authentication accuracy actually stayed the exact same with Hindi but dropped by 7.8% when a local language was spoken. In testing with with 3 different languages combined for the UBM, researchers found that authentication accuracy stayed relatively the same at 82%. In the conclusion, the authors of the paper made an inference on why the SV system returned similar accuracies even though they spoke different languages which was that languages will many phonetic similarities (English and Hindi have around the same phoneme count: 42 and 38 res.) will display the same properties as observed. The researchers also found that creating a UBM with one language, as opposed to multiple ones combined, will result in a higher authentication accuracy, regardless of whether a user is speaking in a different language.<br>Personal Notes:<br>● Although the researchers really only collected audio from one region which all fell under the Tibeto-Burman language family, Arunchal Pradesh is home to over thirty languages with over fifty dialects. Also, many of the people in this area are multilingual.<br>● Different microphones have different sampling rates which is an additional step to consider in the design of a VA system. The addition of this step in the audio db shows that the researchers aimed to create a realistic VA system; not one that works just under experimental conditions.<br>● The researchers chose to use 15-45 seconds of audio data in the validation phase which shows the limits of the Gaussian Mixture modeling. 15 seconds as a min. Is a bit inconvenient |

| | |
|---|---|
| | for users as it would be a much slower method of authentication compared to traditional passwords or fingerprint recognition. |
| Research Question/Problem/ Need | How does speaking in a different language from the language provided at enrollment impact the authentication accuracy in a Speaker Verification System? |
| Important Figures |  |




(b)




(c)

**Figure 1.** These are Detection Error Tradeoff (DET) Curves for the SV system trained with (a) Local (b) Hindi and (c) English and tested with each language. As can be seen, Tani and Hindi have similar tradeoffs in all the plots while English has the lowest False Rejection (FRR) and False Acceptance Rate (FAR) in all the graphs.

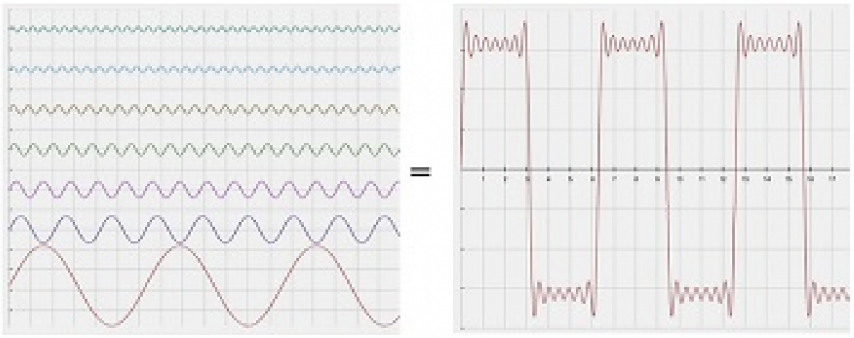
| | |
|---|---|
| VOCAB: (w/definition) | (Many vocab words are already defined in other entries).<br>1) Universal Background Model - A model created to represent general vocal characteristics across an entire population. To |

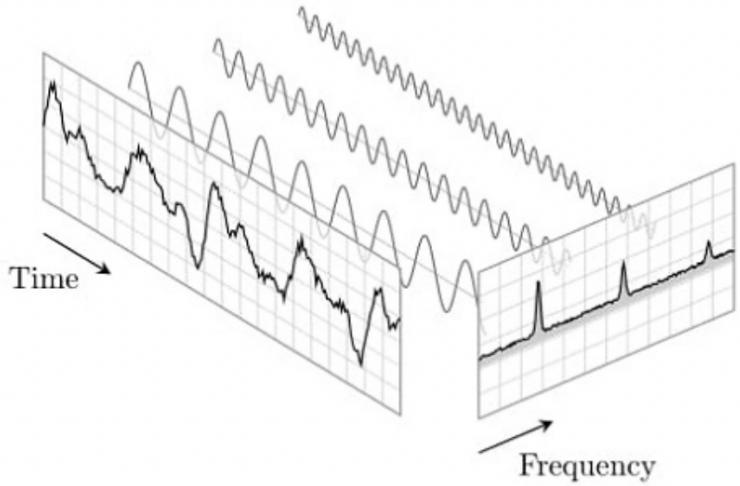
| | |
|---|---|
| | accomplish this goal, UBMs are trained with a large number of speakers. |
| Cited references to follow up on | 1) A. Reynolds, "Robust text-independent speaker identification using Gaussian mixture speaker models", Speech Communications, vol. 17, pp. 91- 108, 1995.<br>2) Kleynhans N.T. and Barnard E., Language dependence in multilingual speaker verification, in Proc. of the 16th Annual Symposium of the Pattern Recognition Association of South Africa, Langebaan, South Africa, pp. 117-122, 2005.<br>3) Haris B.C., Pradhan G., Misra A, Shukla S., Sinha R and Prasanna S.R.M., Multi-variability Speech Database for Robust Speaker Recognition, In Proc. NCC, pp. 1-5, 2011. |
| Follow up Questions | 1) How does the type of microphone used impact accuracy of authentication?<br>2) Would an SV system trained with a language high in phoneme count have a greater accuracy of authentication if a user speaks in a different language with a similar or low phoneme count? (Ex. Training with Vietnamese and using English or French in validation) |

# Article #15 Notes: Audio Pre-Processing For Deep Learning

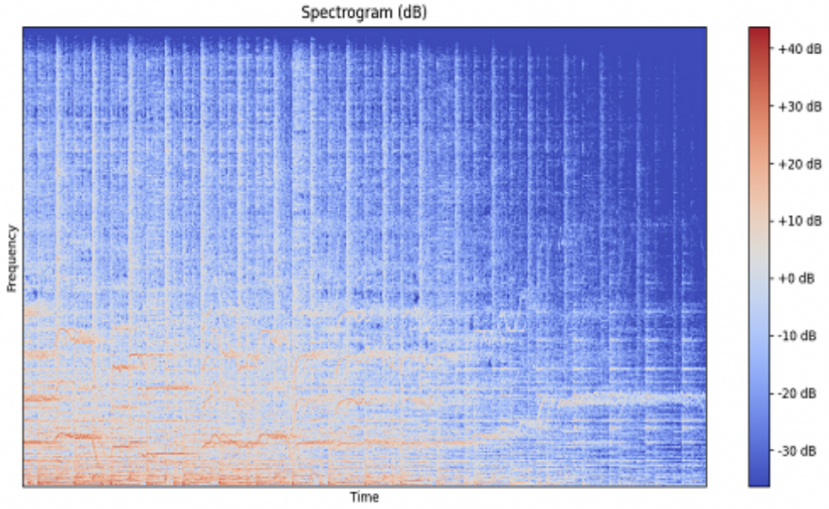
Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Audio Pre-Processing For Deep Learning |
| Source citation (APA Format) | Arzaghi, S. (2020). *Audio Pre-Processing For Deep Learning*. |
| Original URL | https://www.researchgate.net/publication/347356900_Audio_Pre-Processing_For_Deep_Learning |
| Source type | Review Article |
| Keywords | Fourier Transform, Fast Fourier Transform, Short-Time Fourier Transform, Mel Frequency Cepstral Coefficient |
| Summary of key points + notes (include methodology) | This article entitled, "Audio Pre-Processing For Deep Learning," was written by Saman Arzaghi, a researcher at the University of Tehran, in 2020. The main focus of this paper was to articulate how audio is pre-processed or formatted for deep learning. The paper began with a brief introduction of sound. Sound occurs when air is disturbed and the vibrations of that disturbance causes an alternation in pressure. That fluctuation can be graphed as wave in what is known as a waveform. Although sound is analog in nature, meaning that is continuous, it is also prone to change. Since no mathematical function can predict sound, sound must be captured by a computer in digital form or as a series of numbers. This process is known as Analog to Digital Conversion (ADC) and it involves two steps: sampling and quantization. Sampling is the process by a which a computer captures a series of amplitudes in an audio signal at a given rate. Quantization is a compression technique of digital signals by rounding or truncation means. Waveforms plot audio in the time domain but there are implicit features within a wave that cannot be understood in those dimensions, time vs. amplitude. In a process known as Fourier Transformation (FT), a wave can be decomposed into a sum of sin waves oscillating at different frequencies. Fast Fourier Transformation (FFT) is a more efficient algorithm that does the same function and is very important in music analysis. FFTs are needed to compute a Discrete Fourier Transformation which transformations a signal from the time domain into the frequency domain which is frequency vs. signal strength (energy). Although a new domain is gained, energy of different frequencies, the time domain is lost with DFT. Short Time Fourier Transformation (STFT) |

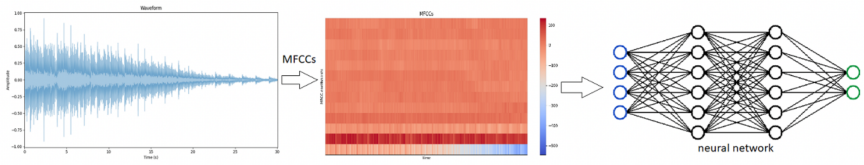
| | |
|---|---|
| | aims to solve this problem by creating a three dimensional graph of time, frequency, and signal strength called a spectrogram. STFT is calculated by taking FFTs at different times and combining them into a spectrogram. Spectograms are important in deep learning because they contain a lot of information about an audio sample that can be understood with a simple waveform. Mel Frequency Cepstral Coefficients (MFCCs) are numbers that measure the timbre of sound. Timbre is an important feature in audio learning as a two audio samples could be at similar frequency and pattern but they could be very different from each other. This phenemon is common in instruments where a piano and violin could be playing the same melody but both have very different musical qualities. MFCCs are an another important input needed for deep learning as they capture only the essential features needed in speech which improve accuracy in classification systems, but also save on computing time. With advancements in deep learning and neural networks, developers don't need to spend as much time with feature engineering, or extracting features from raw data, for analysis as the computer can find those patterns themselves. |
| Research Question/Problem/ Need | How is audio pre-processing important in deep learning? |
| Important Figures |  <br>This *unlabelled* figure shows the process of Fourier Transformation. A complex wave (left) can be decomposed into a series of smaller sin waves (right). |




This figure explains how FFT works in the case of Discrete Fourier Transformation (DFT). A time domain signal is converted into a frequency domain signal.




This plot shows the STFT of an audio sample in a spectogram with color maps. There is a frequency energy of signal strength at every frequency at every time of the sample audio.



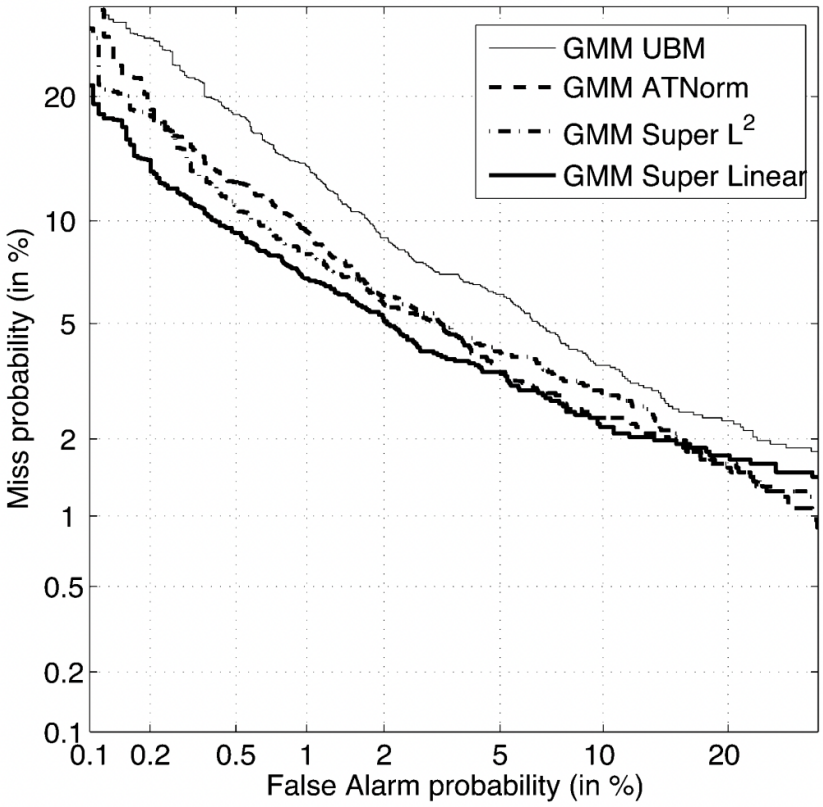

This figure shows the ML pipeline for using MFCCs in neural

| | |
|---|---|
| | networks. |
| VOCAB: (w/definition) | 1) Fourier Transformation (FT) - The decomposition of a complex periodic sound into a sum of sine waves oscillating at different frequencies.<br>2) Fast Fourier Transformation (FFT) - An efficient algorithm for FT calculation. Commonly used in music an audio analysis.<br>3) Discrete Fourier Transformation (DFT) - Using FFT to convert a time domain signal into a frequency domain signal.<br>4) Short Time Fourier Transformation (STFT) - Process of taking the FFT at different times to convert time domain signal into time vs. frequency vs. signal strength plot.<br>5) Spectogram - Graph that shows the STFT of an audio signal. Although this graph is 3d, color maps are commonly used to flatten the image. |
| Cited references to follow up on | Richard G. Lyons. Understanding Digital Signal Processing. 3rd ed. Pearson. 2010. |
| Follow up Questions | 1) How accurate would a ML model that used spectogram as a feature vector as opposed to MFCCs?<br>2) How does the sampling rate affect the training time of a ML model? |

# Article #16 Notes: Support vector machines using GMM supervectors for speaker verification

Article notes should be on separate sheets

| | |
|---|---|
| Source Title | Support vector machines using GMM supervectors for speaker verification |
| Source citation (APA Format) | Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311. https://doi.org/10.1109/LSP.2006.870086 |
| Original URL | https://ieeexplore.ieee.org/abstract/document/1618704?casa_token=H32tV3KSbFMAAAAA:JQz0n8yPh5TZDHWTTN13bhm5otmbDonwrGipd01Qh9jGWuKmeLPVpUDkb0BzfY8IA9CIpuwL |
| Source type | Journal Article |
| Keywords | Gaussian mixture models (GMMs), speaker recognition, support vector machines (SVMs). |
| Summary of key points + notes (include methodology) | This article entitled, “Support vector machines using GMM supervectors for speaker verification,” was written in 2006 for the IEEE Signal Processing Letters publication. This article focuses on whether Support Vector Machines (SVMs) would benefit Speaker Verification systems using Gaussian Mixture Modeling. The goal of the paper was to create an architecture for an SV system using SVMs and test with different speech corpuses or audio databases. SVMs, at their core, try to find a line that best separates two datasets. This is important in Speaker Verification because biometric authentication systems are binary classifiers that compare two datasets, a stored GMM model and a validation GMM model. In this paper, SVMs were used to create different vector norms for comparison of two GMM models. In experimentation, the researchers tested their system with the 2005 NIST speaker recognition (SRE) corpus. For validation, they used a 19-dimensional MFCC vector with a 10ms frame. At enrollment, they used four different types of models: a Universal Background Model (UBM), a GMM using the ATnorm, a GMM using the Super L2 norm, and a GMM using the Super Linear norm. After testing 16078 trials, the results showed that creating a GMM with the Super Linear norm had the lowest Equal |

| | |
|---|---|
| | Error Rate (EER) at 3.77% compared to the EER of a traditional, non-normalized UBM model at 5.68%. |
| Research Question/Problem/ Need | Can using Support Vector Machines (SVMs) improve the authentication accuracy of Gaussian Mixture Model based Speaker Verification systems? |
| Important Figures |  <br>This is a Data Error Tradeoff (DET) plot that compares the False Alarm probability vs. the Miss Probability. Since an optimal system would have an 0% Miss and False Alarm probability, the strength of a system can be determined by the area under the curve. The less area, the more accurate the system is. As can be seen, using a Super Linear GMM results in the lowest area, meaning it also has the lowest Equal Error Rate (EER). |
| VOCAB: (w/definition) | 1) Support Vector Machine (SVM) - Algorithm that finds the hyperplane (ex. Line in 2d feature space) that best separates two datasets<br>2) Equal Error Rate (EER) - An indicator of performance in Biometric Authentication system where a lower % is a better model. It's also used to determine the optimal threshold value in SV systems. Determined by subtracting the FAR and FRR by 100% or adding the TAR and TRR. |

| | |
|---|---|
| Cited references to follow up on | D. A. Reynolds, T. F. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” Dig. Signal Process., vol. 10, no. 1-3, pp. 19–41, 2000. |
| Follow up Questions | 1) How does using a SVM impact the computation time of classification in Speaker Verification systems? |