دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

یادگیری ماشین

فاز اول پروژه

نام و نام خانوادگی
سید محمد جزایری

شماره دانشجویی
810101399

دی 1404

فهرست

چکیده

In this project we will do language classification, but in this phase we are only focusing on data gathering. In this report we will talk about different methods and the challenges of Language Identification.

پرسش 1

- **Definition:** SLID is the process by which a machine determines the language being spoken in an audio clip, regardless of the speaker's identity or the specific words said. It serves as the front-end for multilingual systems like Google Assistant or Alexa, ensuring the correct speech-to-text model is loaded immediately.
- **Controlled vs. Real-World Data:** Unlike laboratory datasets (where speakers read specific scripts in a booth), the audiobooks/podcasts we collected represent in-the-wild data. This includes natural intonation, varied recording equipment, and background noise, making the ML task more challenging but more applicable to real-world scenarios.
- **Closed-set vs. Open-set LID:**
    - **Closed-set:** The model assumes the audio must be one of the N languages it was trained on (e.g., "Is this English, Spanish, or Persian?").
    - **Open-set:** The model must be able to say "I don't know" if it encounters a language it hasn't seen before (e.g., encountering German when only trained on English/Spanish).

پرسش 2

- **Phonetic Similarity:** Languages from the same family (like Spanish and Italian) share similar vowel structures and phonemes, making them harder for a model to distinguish than linguistically distant pairs like English and Mandarin.
- **Speaker & Accent Variability:** A model might accidentally learn the "sound" of a specific narrator's voice rather than the "sound" of the language itself. Accents further complicate this by introducing non-native phonetic patterns.
- **Short Duration Segments:** Our clips are 1-2 minutes long, but models often work with 3-second snippets. The less time a model has, the fewer "rhythmic" and "intonational" patterns it can capture, leading to lower accuracy.

پرسش 3

Why Proper Audio Selection is Critical for Language Identification

The foundation of any Machine Learning system is the quality of its training data, a principle often summarized as **"Garbage In, Garbage Out"**. In the context of Spoken Language Identification (SLID), selection is critical because:

- **Feature Integrity:** The goal is to train models on the inherent linguistic and acoustic properties of a language, such as phonetics, rhythm, and intonation. If a sample contains non-linguistic noise (e.g., heavy background music or digital artifacts), the model may shortcut by learning to identify the noise profile rather than the language itself.
- **Avoiding "Speaker Leakage":** If you select multiple clips of a single narrator with a very distinct or unusual voice, the model might learn to recognize that specific voice

instead of the language. Careful selection ensures the data is representative of the language's broad phonetic structure.
- **Impact on Downstream Phases:** The success of feature extraction (Phase 4) and classification (Phase 5) depends entirely on these initial samples. Preprocessing can clean a signal, but it cannot invent linguistic characteristics that were missing from a poor-quality original recording.

Desirable Properties of Speech Samples for LID Tasks

For a model to successfully discriminate between languages, the audio samples must possess specific, high-quality attributes:

- **Representativeness:** The speech must reflect the standard soul of the target language. This means selecting narrators who use natural linguistic patterns rather than those with exaggerated, unnatural, or highly stylized accents that might skew the model's perception of the language's average phonetics.
- **Clarity and High Signal-to-Noise Ratio (SNR):** The speech should be clear and prominent. While audiobooks are generally clean, avoiding segments with overpowering background music is essential. High SNR ensures that extraction methods like MFCCs capture the vocal tract's shape accurately without interference from external frequencies.
- **Sentence Boundary Integrity:** A desirable sample must begin exactly at the start of a sentence and end exactly after one finishes. This is technically vital because it preserves the prosodic contour (the melody of the sentence). Cutting mid-word or mid-phrase introduces spectral leakage—artificial high-frequency noise created by the abrupt signal cut—which can pollute the feature set.
- **Temporal Sufficiency:** Each clip must be long enough (approx. 1 minute) to provide the model with a sufficient temporal window to observe rhythmic patterns that define a language.

The Role of Listening-Based Qualitative Evaluation

Technical analysis (like checking bitrates or waveforms) is necessary but insufficient. Manual, qualitative evaluation (listening) is the final safeguard.

- **Catching Nuance:** Algorithms are excellent at measuring frequencies but poor at understanding context. A human listener can identify if a narrator has switched to a different language for a quote or if their emotional acting has altered their pitch so much that the sample is no longer representative of the language's normal state.
- **Detecting "Non-Speech" Artifacts:** A human ear can instantly detect subtle issues—like a narrator's heavy breathing, page-turning sounds, or minor digital stutters—that might appear as valid data points to a feature extraction algorithm but would actually degrade a model's performance.
- **Ensuring Naturalism:** Qualitative evaluation confirms the audio reflects natural linguistic patterns. This ensures the model learns the actual language rather than the specific mechanical quirks of an audiobook's production.

پرسش 4

1. Fast Fourier Transform (FFT)

- **How it works:** FFT decomposes a signal from the Time Domain (amplitude over time) into the Frequency Domain (amplitude over frequency). It identifies which pitches are present in a short window of speech.
- **Suitability for LID:** Low. Raw FFT is too detailed and sensitive to noise/pitch; it captures the speaker's voice more than the language's structure.
- **Pros/Cons:**
    - Pro: Computationally very fast.
    - Con: Lacks perceptual scaling; it treats high frequencies (which humans hear poorly) the same as low frequencies.

2. Mel-Frequency Cepstral Coefficients (MFCCs)

- **How it works:** MFCCs apply a Mel Scale to the FFT. The Mel scale mimics human hearing by spacing out frequencies—we are better at telling apart low-pitched sounds than high-pitched ones. It then uses a Discrete Cosine Transform (DCT) to compress the information.
- **Suitability for LID:** Very High. It captures the shape of the vocal tract, which is the primary way languages differ phonetically.
- **Pros/Cons:**
    - Pro: State-of-the-art for speech recognition; robust against speaker-specific pitch.
    - Con: Sensitive to loud background noise (though less of an issue for our clean audiobooks).

3. Log-Mel Spectrogram

- **How it works:** This is a visual representation of the spectrum of frequencies as they vary with time, with a logarithmic scale applied to the magnitude.
- **Suitability for LID:** High. It allows Deep Learning models (like CNNs) to see the patterns of a language as if it were an image.
- **Pros/Cons:**
    - Pro: Retains more temporal information than MFCCs.
    - Con: Requires more storage and processing power.

4. Zero-Crossing Rate (ZCR) & Spectral Centroid

- **How it works:** ZCR measures how often the signal switches from positive to negative (percussive/noisy sounds have high ZCR). Spectral Centroid marks the center of mass of the sound (the brightness).
- **Suitability for LID:** Moderate. Helpful for distinguishing fricative-heavy languages (like English or Persian) from vowel-heavy ones.
- **Pros/Cons:**
    - Pro: Extremely simple to calculate; good for rhythm analysis.
    - Con: Too simple to be used alone; must be combined with other features.

پرسش 5

- **Definition:** Instead of a model saying "This is 100% Spanish," Similarity Learning (or Metric Learning) teaches the model to calculate a distance between two audio samples. If the distance is small, the languages are likely the same.
- **Loss Functions:**
    1. **Contrastive Loss:** Uses pairs. It minimizes the distance between Positive Pairs (two English clips) and maximizes the distance between Negative Pairs (English vs. Persian) until they are separated by a specific margin.
    2. **Triplet Loss:** Uses three inputs at once: an Anchor (the reference), a Positive (same language), and a Negative (different language). The goal is to push the Negative away while pulling the Positive closer to the Anchor simultaneously.
- **Application to Language Discrimination:** This is vital for Open-set LID. If you provide the model with a language it has never heard, a standard classifier would guess English or Persian. A Similarity Learning model would see that the German clip is far away from both English and Persian in the embedding space and correctly identify it as Unknown.