nexidia

WHITE PAPER

Automatic Language Identification of Audio

A White Paper by Nexidia, Inc.

Abstract

There are a numerous monitoring applications where a large body of audio data is collected without knowing a-priori what languages may be spoken in the recordings, or which language is being used in any specific recording. For any reasonable sized collection, it is wholly impractical to have trained analysts listen to each call to identify which language is spoken, so the call may be routed for analysis appropriate to that language.

To address this problem, Nexidia has developed a set of tools that completely automate the process of analyzing an audio clip to determine with high accuracy what language is being spoken within that clip. These tools provide rapid identification (currently processing at up to 100 times faster than real-time speeds) combined with the flexibility of allowing a user to custom train models for a specific audio environment and set of languages.

In addition, the Nexidia tools can identify multiple languages within the same audio clip and adapt to "non-spoken" languages such as distinguishing music from speech.

Components

The Nexidia language identification (LID) workbench consists of two major components:

- A set of C++ and COM libraries providing access to the core language identification APIs.
- A training tool application for creating language identification models of 2 or more languages of interest

Training a LID model

In order to perform language identification, you must first train a model for the languages of interest in your particular environment. As part of the language identification workbench, Nexidia provides an application that creates a model based on sample audio for each language you wish to identify. In general, at least 20 hours of each target language should be collected (preferably more if it is available). A single language identification model can be developed for any number of simultaneous languages (up to a practical limit of approximately 40 languages depending on the run-time system hardware configuration).

In addition to the audio data, model creation requires the use of one or more Nexidia supplied language reference discriminators. The discriminators provide a baseline for comparing the languages during the training process. The number of discriminators you use directly influences the run-time speed of identification and the overall accuracy of language identification. The more discriminators you use, the more accurate the model will be, but the runtime speed will also be slower, as summarized in Table 1.

NUMBER OF DISCRIMINATORS	RUN-TIME SPEED (TIMES FASTER THAN REAL-TIME)
1	100
3	25
4	18
8	8
13	6

Table 1 – Reference models effect on run-time identification speed

Nexidia provides several sample models trained on data provided by the Language Data Consortium . While these models are often suitable for demonstration and preliminary evaluation, they cover only a few languages and represent a narrow range of acoustic environments. Nexidia has found that training on representative data from the customer environment always produces the best results.

Simple API calls

Once you have a model, the LID workbench provides a set of API calls to use the LID model to identify the language for an audio clip. Similar to the Nexidia phonetic search framework, you can work directly with audio streams or files using either C++ or COM interfaces. The output of Nexidia LID is a score in the range of 0.0 – 1.0 for each language in the model.

LID scores are always relative to the other languages in the language model. That is, each score represents how much more likely an audio clip matches one language than another. For instance, using a three language model built for English, Spanish, and German, a score of 0.9 for English would mean that there is approximately a 90% chance of the audio being English. If the clip were none of the 3 languages, you might still get a high score for English as the clip was a much better match to English than either Spanish or German.

Nexidia provides a number of additional options to support a broad range of applications, including:

- Can identify any portion of an audio file in any of the formats supported by Nexidia
- Retrieve the highest scoring language and its associated score
- Retrieve the scores for all of the languages in the model for the audio clip
- Identify transitions from one language to another within an audio clip

Being able to identify transitions between languages is quite powerful. This features lets you work with audio where the language shifts, and to identify the exact location where a different language is spoken. As an example, Nexidia uses this feature to separate speech from music within a recording in order to eliminate erroneous results within songs or advertisements.

Through the API, an application can control how often transitions are allowed so that, for instance, clips are marked as continuous speech in the presence of brief tones (such as DTMF tones) rather than as rapidly alternating music and speech. Similarly, an application might not want to mark a clip as another language simply because it contained a brief phrase (descriptive foreign words or place names).

Integration with Nexidia phonetic search framework

Language identification is designed to work directly with the Nexidia phonetic search framework. Generally; LID is used as a front-end system to identify incoming files. Once identified, the files can then be indexed with the appropriate language pack and made available to the users who work with that specific language. Because of the inherent speed and scalability of Nexidia language identification, audio analysis can keep pace with very large scale continuous data collections across multiple languages.

NUMBER OF DISCRIMINATORS	MIN CDET	SPEED (TIMES REAL-TIME)
1	8.9%	102
2	8.4%	39
3	7.1%	25
4	7.1%	18
6	6.3%	11
8	6.5%	8
13	5.3%	6

Table 2

Accuracy – NIST 2005 LRE test

LANGUAGE IDENTIFICATION

The Nexidia Language Identification system was evaluated on the NIST 2005 Language Recognition Evaluation test set. The primary condition contained seven languages: English, Hindi, Japanese, Korean, Mandarin, Spanish and Tamil. The English contained two dialects, that spoken in North America and that spoken in India (Hindi-accented English). The mandarin consisted of Mainland and Taiwan dialects. Each file had an average of 30 seconds of voice. There were a total of 2421 files in the test set. The accuracy is reported as the minimum CDet as defined by the NIST accuracy criteria.

The final results are shown in Table 2 for several models. The models shown vary from 1 to 13 language discriminators. As mentioned previously, more discriminators give higher accuracy, but processing speed decreases proportionally.

As can be seen from the table, one discriminator gives a minimum CDet of 8.9% at a processing speed of 102 times faster than real-time on a 3.2 GHz Xeon. The most accurate model had a 5.3% minimum CDet with a processing speed of 6 times faster then real-time. Of particular note is that for Nexidia language identification, run-time processing speed is independent of the number of languages in the model.

DIALECT DETECTION

The results for the NIST 2005 LRE dialect tests are shown in tables 3 and 4. Table 3 shows the CDet on the English dialect test for both the 1 and 8 discriminator systems. The two dialects were North American and that spoken in India (Hindi-accented). Table 4 shows results for the Mandarin dialect test.

NUMBER OF DISCRIMINATORS	MIN CDET	SPEED (TIMES REAL-TIME)
1	13.4%	102
8	9.7%	8

Table 3 – Accuracy for Hindi-accented English

NUMBER OF DISCRIMINATORS	MIN CDET	SPEED (TIMES REAL-TIME)
1	27.6%	102
8	20.7%	8

Table 4 – Accuracy for Mandarin dialects

Dialect detection can be much more demanding than language detection. For dialects where there are significantly different speech patterns, Nexidia LID performs well. When the differences are primarily due to choices of words, vocabulary, or where the pronunciations are on slightly different on a subset of words, then it will not perform as well, as shown by the difference in performance between Mandarin dialect detection and English dialect detection in the two tables.

Processing performance

Nexidia language identification is not only flexible and accurate, it is also quite fast. As shown previously in Table 1, speeds up to 100 times faster than real-time on a single 3.2Ghz Xeon processor are obtainable using a single language discriminator. For example, a one-hour audio file can be processed in 36 seconds. For many applications, it is not necessary to process the entire media file to determine the language. Instead, the first 1 or 2 minutes is sufficient to identify the entire recording. In those situations, overall processing throughput can be greatly increased since only a fraction of the audio must be processed to identify the language of the recording.

A unique feature of Nexidia LID is that the speed of identification is virtually independent of the number of languages in the language model. The difference in throughput between a 2 language model and 20 languages is less than 5%. There is currently a memory limit on the size of a LID model that effectively restricts the size of a single model to support up to 50 simultaneous languages. Other than this limitation, system throughput is not significantly affected by the number of languages being detected.

Adapting to generalized languages (music identification)

Nexidia Language Identification generalizes the concept of a language to allow for discrimination between, for example, music and speech. We define one "language" with a large collection of music, including examples of instrumental and vocal music across a wide variety of formats. The other is a collection of purely spoken English, with no music in the background. We then trained a model using these two audio sets. The results are quite good at identifying music within broadcast audio, and not quite as effective for telephony band audio.

Combined with ability to detect transitions between languages, music identification is especially useful for filtering out unwanted results in an analysis. For instance, searching for stories on "car insurance", you would probably want to filter out commercials containing that phrase. For broadcast, these can almost always be separated from news stories by the presence or absence of background music.

Nexidia language identification also provides control over how to decide which language to choose when both are spoken at the same time, for instance, speech against a musical backdrop. The application can set thresholds to mark a clip as music even if there is only very low volume music in the background. Or you may choose to mark it as speech except there is very loud (relative to the speech) music.

Current LID research topics

In addition to the current commercially available language identification product, Nexidia is also working on a number of research areas to improve the accuracy and utility of the language identification suite. Current features under research are:

- Provide an ability to add a new language to an existing model without having to retrain the entire model.
- Improved dialect detection
- Provide a mechanism to define a "catch-all" category to identify out-of-language (audio that does not match any of the languages in the model)
- Automatically find portions of the audio with speech in order to improve accuracy by ignoring silence, line noise, and other non-speech signals.

Summary

In this paper, we have described a suite of tools for automating the process of identifying the spoken language in audio documents through the Nexidia Language Identification workbench. Nexidia LID provides an extremely fast, robust, and flexible system which can be adapted to a wide range of applications when the source of audio content is either unknown, or too complex to infer by manual means. The supplied LID training tool gives the customer a unique ability to tailor the identification models to their specific set of languages and acoustic environment.

Copyright Notice