LDC: DATA TYPES

Broadcast Narrow Band Speech

A New Data Type

Broadcast Narrowband Speech (BNBS) is a new data type used in the development and evaluation of Language Recognition (LR) technologies. The sources of BNBS are narrowband segments embedded in broadcast, typically listener call-ins, phone interviews of pundits and some correspondent reports and man on the street interviews. These data address a need that is otherwise difficult or expensive to fulfill, that of numerous short segments of many different speakers in many different languages.

BNBS was the principal source of data in the 2009 Language Recognition Evaluation (LRE) organized by the U.S. National Institute for Standards and Technology (NIST).

BNBS Features

As a data source, BNBS meets or exceeds most of the requirements for speech data that supports LR technologies.

Such data should be abundant, providing one sample from each of hundreds speakers in each of dozens of languages. Each speech segment should contain speech from just one speaker in just one language. The segments should be at least 30 seconds in duration and there should be no correlation between language and channel conditions. Specifically it should not be that case that all speakers of a given language were recorded in a way that is unique to that language.

Since BNBS is extracted from broadcast and webcast audio there are large archives of radio, television and web programming that contain it. Several broadcast sources provide programming in dozens of languages and although the incidence of BNBS varies by program, it is plentiful on average. Finally, since the speakers may be calling from landlines or cell phones in different countries where the target language is spoken and since the call is then broadcast, the signal characteristics of the original calls are conflated with those of the broadcast reducing any correlation between signal and language features.

In comparison with Conversational Telephone Speech (CTS), the data type more commonly used in NIST LRE campaigns, BNBS is more plentiful and cost-efficient to collect in a larger range of languages. Voice of America alone broadcasts in 47 languages in 2009 that the service labels:

Afan Oromo	Albanian	Amharic
Armenian	Azeri	Bangla
Bosnian	Burmese	Cantonese
Chinese	Creole	Croatian
Dari	English	French
Georgian	Greek	Hausa
Hindi	Indonesian	Khmer
Kinyarwanda	Kirundi	Korean
Kurdi	Kurdish	Lao
Macedonian	Mandarin	Ndebele
Pashto	Persian	Portuguese
Russian	Serbian	Shona
Somali	Spanish	Swahili
Thai	Tibetan	Tigrigna
Turkish	Ukrainian	Urdu
Uzbek	Vietnamese	

Disadvantages of BNBS result from lack of control of the conditions in which the data are created. The language of the narrowband segment may differ from that of the remainder of the program and there may be multiple recordings of the same speaker that are not recognized as such. However, these disadvantages can be managed with efficient human auditing or overwhelmed by large volumes of data.

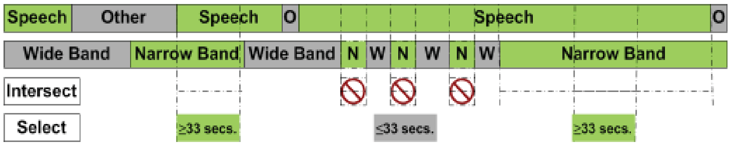
Availability

BNBS data were used in NIST's 2009 Language Recognition Evaluation (LRE) and will be used for subsequent evaluations. Researchers may acquire BNBS by joining the next NIST evaluation, planned for 2011, or by licensing it from LDC. The first BNBS corpus will appear in LDC's Catalog in 2010. Please refer to NIST's 2009 LRE website for more information: www.itl.nist.gov/iad/mig/tests/lre/2009/

BNBS Collection

LDC collects BNBS by first identifying sources that create programming in multiple languages. This is not only efficient but it also reduces the connection between linguistic and channel features that may hinder the development and evaluation of LR technologies. Once sources are identified, LDC records entire programs in languages of interest. Currently, all languages are considered interesting though future need may focus attention on individual languages that are under-represented or on sets of languages that are difficult to distinguish from each other.

The next step is to identify regions of speech that are narrowband. This is done in multiple stages. First, the audio is passed through a speech activity detection system to eliminate from further consideration any silence, music or, in theory, any other non-speech signal. A second filter distinguishes any narrowband signal. From the intersection of these filters, segments greater than 33 seconds are selected from which the initial and final 1.5 seconds are trimmed.



Audit

Because BNBS is found data, LDC audits to establish the ground truth critical for technology evaluation. The audit determines whether the segments include only speech in the target language, confirms that they are narrowband and identifies the number of speakers and whether they have been heard previously.

Auditors are fluent speakers of the target language -- but not necessarily of each dialect represented -- who are trained and tested in narrowband detection. Auditors are presented with 30-second segments, one at a time, and listen to each in its entirety using high quality headphones.

For each segment, auditors indicate whether it consists entirely of speech entirely in the target language, is telephone-like in quality and comes from a single speaker who has not been heard previously. Auditors also label the segment for sex of the speaker, quality of the audio and whether the speech seems native and standard, heavily regional or non-native.

Language versus Dialect

BNBS may originate with multiple sources that label linguistic varieties differently. One source may treat two varieties as dialects of the same language while another may treat them as separate languages, perhaps devoting different programming to each. The BNBS collected so far at LDC comes from a single source, Voice of America. LDC has chosen to respect the names and statuses of linguistic varieties asserted by their broadcast sources while simultaneously providing the alternate names used for each and indicating which pairs are mutually intelligible. Auditors also indicate if any segment contains more than one speaker or any speech in a language other than the target.

Formats

As found data, BNBS naturally varies widely in audio format. Channels, sampling rate, sample size, compression and audio file headers vary independently in complex ways. To support LR technology evaluation, LDC normalizes evaluation data, prior to distribution, to flac compressed, linear sampled audio at 8kHz sampling rate with 8 bit samples and NIST SPHERE file headers. This format most nearly approximates the format used for prior NIST evaluations though audio may have been compressed with loss prior to LDC collection.

Metadata

The metadata for each audited segment, based on human audit, includes the following:

- unique segment identifier
- relative path to corresponding audio
- corpus from which the audio is drawn
- offset in seconds from beginning of audio corresponding to where segment begins/ends
- language spoken in segment plus optional comments
- sex of speaker
- categorization of speech as native, non-native or regional
- whether single or multiple speakers in segment
- whether speaker already present in corpus
- quality of the signal with optional comments
- whether the segment is narrowband
- whether segment is composed entirely of speech