IARPA Babel Data Specifications for Performers

Contents

1 Introduction

This document provides specifications of the conversational and scripted telephone speech data collected using mobile and fixed telephone networks, as part of the IARPA Babel program.

It also includes information related to data delivery for Babel performers. Specifications related to data structure are provided along with information related to the content of the build pack.

2 Database Contents

2.1 Contents for Scripted Portion

Scripted data items are detailed in the following table:

Corpus identifier	Item identifier	Num. Items per speaker	Corpus contents		All languages but Cantonese	Cantonese only
A	1	1	terms of address		✓	
C	1	1	number (5+ digits)	connected digits	✓	✓
C	2, 5, 6	3	telephone number (9-11 digits) (2 read, 1 spontaneous)			
C	3	1	credit card number (14-16 digits)			
C	4	1	PIN code (6 digits) (set of 150 codes)			
D	1	1	spontaneous date, e.g., birthday	dates: 4 for Cantonese, 5 for others	✓	✓
D	2, 4	2	prompted date, word style i.e., not digital		✓	✓
D	3	1	relative and general date expression			✓
D	3, 5	2	relative and general date expression		✓	
I	1-2	2	isolated digits		✓	✓
L	1	1	spelling of proper name, spontaneous (e.g., own forename), see O1	3 spelled words (letter sequences)		✓
L	2	1	spelling of directory assistance city name (see O3)			
L	3	1	real/artificial to maximize letter coverage			
M	1	1	money amount in local currency, mixed size and units		✓	✓
N	1	1	natural number		✓	✓
O	1	1	proper name, spontaneous (e.g., own forename)	directory assistance:		✓

O	1, 8	2	proper name, spontaneous (e.g., own forename)	7 names for Cantonese, 8 for others	✓	
O	2	1	city of birth / growing up (spontaneous)		✓	✓
O	3, 4	2	most frequent cities (set of 500)			
O	5	1	most frequent companies/agencies (set of 500)			
O	6, 7	2	"forename surname" (set of 150 "full" names)			
S	0-9, A, B, C	13	phonetically rich sentences		✓	✓
T	1	1	time of day (spontaneous)	3 time phrases	✓	✓
T	2, 3	2	prompted time phrase, word style i.e., not digital			
R	1	1	the "silence word" recording		✓	✓

Table 2-1 Corpus contents definition

2.1.1 Isolated digits (I1-I2)

Two isolated digit utterances are elicited for all databases. This is READ (or PROMPTED in answer to a simple question).

2.1.2 Digit/number strings (C1-C6)

Three/four continuous digit or number strings are elicited by presenting DIGIT sequences, not words, in order to elicit the natural spoken forms of read numeric expressions.

2.1.2.1 Number (C1)

A unique number is provided to elicit READ connected digits rather than fluent numbers.

2.1.2.2 Telephone number (C2, C5, C6)

Three types of telephone numbers are elicited: two READ and one spontaneous.

Read (C2, C5): the two 9-11 digit READ telephone numbers are elicited using presentation that reflected typical telephone numbers for national numbers including area codes, but not PBX extensions. For example:

0171 343 8979 (UK telephone)
1.23.46.88.99 (FR telephone)

Spontaneous (C6): each speaker pronounces one telephone number. To preserve anonymity, the telephone number will be the answer to:

Please, say the telephone number of a friend.

2.1.2.3 Credit card number (C3)

This is a READ 14-16 digit credit card number. The number, spacing and presentation of digits should reflect what is printed on a range of typical credit cards, such as VISA/MasterCard and Amex formats.

XXXX XXXX XXXX XXXC	(16-digit VISA/MasterCard/JCB/Discovers card)
XXXX XXXXXX XXXXC	(15-digit American Express)
XXXXX XXXX XXXXC	(14-digit Diners and Carte Blanche cards)
XXX XXX XXXX XXXC	(typical 14-digit telephone calling card)

2.1.2.4 PIN code (C4)

The PIN code is a READ connected digit string of length 6.

2.1.3 Natural number (N1)

At least one natural number phrase is elicited.

2.1.4 Money amounts (M1)

Prompts are provided to elicit typical phrases used with money amounts, including currency words for the native language. Items are READ. There are a mixture of small (i.e., including decimal currency units) and larger money amounts (not including decimal currency units), although this is country-dependent.

2.1.5 Times (T1-3)

Times are normally said in either a DIGITAL format, e.g., "*23:49*" or ANALOG format, e.g., "*half past seven*". Two time phrases are elicited:

- one SPONTANEOUS phrase (T1) is elicited by asking the caller for the current time of day. The caller has full freedom whether to respond in DIGITAL or ANALOG form, and to his/her choice of degree of precision,
- two READ phrases (T2-3) are in ANALOG form. , including equivalents of:

  AM, PM, half, quarter, past, to, noon, midnight, morning, afternoon, evening, night, today, yesterday, tomorrow, minutes, hours, o'clock

2.1.6 Dates (D1-5)

There are two ways to specify a date: DIGITAL, e.g., "*27/12/96*", or ANALOG, e.g., "*Vendredi, premier Mai 2003"*. The focus of the collection is on ANALOG forms. This includes:

weekdays, months, ordinal numbers, year expressions

One SPONTANEOUS date (D1) is elicited, e.g., the speaker's birthday. Two READ dates (D2, D4) are elicited in ANALOG form, covering all weekdays and months, e.g., "*Monday, the first of May 2003*; *May the first, 2003"*.
The other READ dates (D3 for Cantonese; D3, D5 for other languages) include RELATIVE and general date expressions, including:

today, tomorrow, the day after tomorrow, the next day, the day after that, next week, Good Friday, Easter Monday, etc.

2.1.7 Spelled words (L1-3 for Cantonese only)

The spellings are READ (L2 and L3), except for one SPONTANEOUS forename spelling (L1).

2.1.8 Phonetically rich sentences (S0-9, A, B, C)

These sentences were READ from the prompt sheets. These phonetically rich sentences were provided to obtain adequate training coverage of monophones and most frequent biphones and triphones for continuous speech modelling using subwords. It is not intended to provide phonetically balanced material, where the frequencies of occurrence of the phones and contexts mirror that of "*typical*" linguistic material in that language.

2.1.9 Directory assistance names (O1-O7 for Cantonese, O1-O8 for other languages)

These items provide a useful evaluation set for directory assistance applications, and are defined as:

- 1 spontaneous proper name (O1) for Cantonese; 2 different spontaneous proper names (O1) for other languages;
- 1 spontaneous city/town name (O2);
- 2 items drawn from a set of approx. 500 most frequent cities (O3-4);
- 1 item drawn from a set of approx. 500 most frequent companies/agencies (O5);
- 2 proper names drawn from a set of approx. 150 (forename-surname spoken together) (O6-7).

2.1.10 Terms of Address (A1 for all languages but Cantonese)

One item was prompted out of a closed set of words used to address people during conversation. The list could include terms of address in as many contexts as possible out of *inferior speaking to superior, superior speaking to inferior, and equal speaking to equal*, and in different social/formal situations, such as *between family members, friends/close associates, and strangers/formal associates*. The closed set of terms could include, for example, terms of address used to address people in the judicial system, in a religious context, military ranks (if used when addressing a person directly), etc. The number of terms of address in this closed set may vary depending on the language, and ranges between a minimum of 10 and a maximum of 50, depending on the richness of terms in a given language. Examples of typical terms of address are:

Sir, mr., reverend, your honor, bro, dude, lord, doctor, captain

2.1.11 Silence word (R1)

Background noise is recorded. The speaker is prompted not to talk during this recording. The recording has a length of 10 sec.

2.2 Contents for Conversational Portion

Conversational data are made up of free-speech, natural and fluent conversation between two speakers in a given language. The only guidance given to speakers in relation to the content of the conversation is to speak on a topic of their choice from a very general list of topics to ensure broad coverage of topics/vocabulary across the database while at the same time keeping the subject general, to encourage good vocabulary coverage and natural conversations. See list of topics in Section 3.3. Speakers were also directed to avoid discussing the recording of the conversation or any instructions given in relation to the recording.

The two speakers in each conversational call talk for approximately 10 minutes. Should a call drop out for any particular reason (such as bad mobile coverage, low battery etc.), the speakers are allowed to reconnect and complete the call until they have spoken for 10 minutes, thus a recording session may comprise more than one recording per speaker. Short call attempts with no actual content are discarded and not included in the database.

3 Recording Standards

3.1 Recording Platform

Calls are recorded from an ISDN connection with the terrestrial telephone network, i.e., 8 kHz sampling rate, A-law or mu-law coding or an equivalent digital interface. In conversational data, each of the two speakers is recorded on a separate channel which is stored on a separate signal file.

Recording Specifications:

- Care is taken to ensure that in general each utterance contains approximately 200 ms. or more of silence at the beginning and 200 ms. or more of silence at the end (guaranteeing that the speech is not truncated.).
- Recordings should not contain big variation in speech sound volume.
- A recording may be rejected if the speaker articulates poorly or mumbles throughout the call only if this happens to an extent that the speaker is generally not well understood.
- Repeated fragments, reformulations, misspeaks, hesitations are acceptable.

Speech signal requirements:

- No typical saturation of the speech signal – slight saturation characteristic of using cellular telephone calls to a remote platform as a recording medium is acceptable.
- The speech signal will not be modified by any pre or post processing steps, except the conversion of the recorded digital information to the specified delivery format. Files with signal amplification, filtering, and noise reduction alterations are not acceptable.

3.2 Scripted Data

Prompt sheets are used and are distributed to speakers. The prompts consist of text to be read (generating read speech) and questions or tasks to be answered (generating short spontaneous speech).

3.3 Conversational Data

The speakers are encouraged to talk about one of the topics in the following list of topics that they feel most comfortable discussing:

- Family
- Friends
- News & Current Affairs
- Culture
- Sports
- Work
- Holiday / Leisure Miscellaneous
- Study/School
- Health
- Movies/TV Shows
- Information and Technology

However, speakers are not limited to these topics and are permitted to speak on other topics they feel comfortable discussing.

Since conversational speech is recorded via the telephone it is not certain that speakers can be instructed personally. If this is the case, an instruction sheet is provided.

4 Environmental and speaker specific coverage (both Scripted and Conversational)

4.1 Speaker Specific Coverage

Speakers are recruited to provide broad coverage of age, gender, and dialect for each data collection.

4.1.1 General constraints

The caller should be a **native** speaker except for those databases where a second language or a *lingua franca* is recorded. In these cases, the speaker should be a **fluent** speaker in such language.

4.1.2 Gender balance

The database aims to provide a gender balanced data set, but this may not be practical in some data collection contexts.

4.1.3 Age distribution

For some languages, and for some dialects within a given language, comprehensive coverage of speakers in the older range may not always be possible. Callers younger than 16 or above 60 are optional.

4.1.4 Dialectal regions

Dialect regions are defined for each collection language in cooperation with IARPA, and speakers are recruited to provide coverage of these dialects. The number of dialects to be collected for a given language/region is typically in the range 3-6, and does not normally exceed 6.

4.2 Environmental Coverage

Though there can be many types of environmental variations, only those relevant are covered. The acoustic environment conditions and network specifications are described below.

4.2.1 Acoustic environmental conditions

There are many possible environments from which a mobile caller may conceivably make their call. The following 5 acoustic conditions have been chosen as representative of a mobile user's environment:

- Passenger in moving vehicle, such as car, railway, bus, etc. (background traffic "*emission noise*", combined randomly)
- Public place, such as bar, restaurant, etc. (background talking)
- Stationary pedestrian by road side (background traffic "*emission noise*")
- Quiet location, such as home, office, ... (quiet environment)
- Passenger in moving car using a hands-free car kit for mobile phones

4.2.2 Network specifications

In many countries several network providers offer mobile and/or fixed line telephone services. To exclude network provider specific peculiarities, an attempt is made to cover the different providers. The identity of the provider is documented for each call if available. There are no constraints on the distribution of calls per network system.

4.2.3 Telephone model

Steps are taken to promote telephony handset diversity. The type of handset is documented for each call if available.

5 Transcription Conventions – Telephony Speech Database

5.1 Introduction

The transcription is intended to be an ORTHOGRAPHIC, lexical transcription with a few details included that represent AUDIBLE ACOUSTIC EVENTS (speech and non-speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance.

The transcription is intended to be a quick and broad transcription. Transcribers were told not to agonize over decisions. The overall aim is to keep as much speech in the corpus as possible and to avoid the need for deleting recordings from the corpus due to some extra noises, disfluencies, etc.

The character set to be used for the orthographic transcriptions is UTF-8. A Language Specific Peculiarities Document (LSP) describes the format of the orthographic transcription for each of the languages to be collected. Additionally, the document provides specific information on languages that do not use white space as word boundaries and/or do not use a Latin alphabet. In such cases, the LSP contains a description of the method used for word boundary detection as well as a complete list of the symbols used in Romanization and their mapping to the phone alphabet used for the phonetic transcription of the particular language.

5.2 Speech Events

5.2.1 Hesitancies

Hesitancies (fillers) are sounds made by the speaker to indicate that they are still continuing but have made a mistake or are thinking of what to say next. In English, these would include fillers such as “ah”, “um”, “er”, “hm”, etc. These are all tagged using the <hes> tag. For each language, a list of possible fillers is documented and corresponding pronunciations are provided in the pronunciation lexicon.

5.2.2 Mispronunciations

If the speaker mispronounces a word, the word is spelled correctly and annotated using the * * tag.
For example:
“representive” -> *representative*

5.2.3 Fragments

If a speaker stumbles mid-word it is transcribed up to the cut-off point and hyphenated.
For example:
"to- tomorrow" -> to- tomorrow
For example, if the word is not repeated correctly:
"to- the day after tomorrow"-> to- the day after tomorrow

In less frequent cases the hyphen will occur at the beginning of the word. For example: -morrow

5.2.4 Unintelligible Speech

If a word is unintelligible the (()) tag is utilized. The unintelligible tag is used regardless of whether it is a single word or a string of speech.

5.2.5 Truncations

Truncations may occur at the very beginning or end of an utterance if the recording device has cut off a word. The affected word shall be transcribed in full and marked with the ~ tag.
For example:

~ satisfactory	(truncation at beginning of utterance)
unsuitable ~	(truncation at end of utterance)
(()) ~	(truncation with unintelligible word)

While a truncation is similar to a fragment, it is not marked as a fragment.

5.2.6 Foreign Words and Code-Switching

The <foreign> tag is used when the speaker says a word or string of words from another language that would not be widely accepted or understood as part of the native language. This utterance is NOT transcribed and the <foreign> tag is inserted instead.

Individual loan words that are spoken and commonly used as part of the native language are transcribed with the accepted loan word spelling. For example, words such as “kimono”, “croissant”, or “falafel” would be considered commonly accepted loan words in the English language. Such words are written using the same character set as the native language.

Please note that speakers are directed to only speak in the language of the collection. Any calls with excessive amounts of foreign language use or code-switching are rejected.

5.3 Non-Speech ("Acoustic") Events

These are categorized as either foreground or background sounds.

5.3.1 Background sound

Continuous low level background noises do not need to be tagged. For continuous medium to loud background noise, the <sta> tag is inserted once at the point where the sound begins.

5.3.2 Foreground sounds

Events are only transcribed if they are clearly distinguishable. Very low-level, i.e. non-intrusive events are ignored. The event is transcribed at the place of occurrence, using the defined symbols in angle brackets. For noise events that occur over a span of one or more words, the transcription should indicate the beginning of the noise, just before the first word it affects. If a noise occurs more than once in sequence, the appropriate tag is only be inserted once.

The first four categories of acoustic events originate from the speaker, and the other categories originate from another source. Sounds originating from the speaker usually do not overlap with the target speech, and sounds originating from other sources can of course occur simultaneously with the speech.

The categories are:

<lipsmack>	lip smacks, tongue clicks
<breath>	inhalation and exhalation between words, yawning
<cough>	coughing, throat clearing, sneezing
<laugh>	laughing, chuckling
<click>	machine or phone click
<ring>	telephone ring
<dtmf>	noise made by pressing telephone keypad
<int>	any other intermittent foreground noise

If any of the above events overlap with a word and the event is loud enough to render the word useless, the appropriate tag is inserted and the word is not transcribed.

5.3.3 No Speech

Even if there are some foreground sounds, only the <no-speech> tag is used if there is no actual speech for more than one second. For example:

"silence <breath> silence" should be transcribed as "<no-speech> "

5.3.4 Overlap

Overlap occurs when two foreground speakers talk at the same time. This utterance is NOT transcribed and the <overlap> tag is inserted instead.

5.3.5 Prompt

This tag is used for an electronic voice or automated recording. This utterance is NOT transcribed and the <prompt> tag is inserted instead.

5.3.6 Change of Speaker (for different gender only)

Demographics of speakers are specified in the first utterance of each conversational session. The following tags are used to indicate if the gender of a speaker changes during a call. The appropriate tag is inserted at the exact point at which the new speaker starts:
<male-to-female>
<female-to-male>

5.4 Spelling

5.4.1 Proper Names

Proper names are transcribed in a case-sensitive manner in applicable languages. Initials should be in capital letters with no period following. For example:

“George W Bush has confirmed his relationship with the South American government”.

5.4.2 Titles and abbreviations

All titles and abbreviations are transcribed as a word. For example:

Dr -> Doctor

Exception: if the abbreviated form is actually pronounced. If speaker says 'Appen Butler Hill Inc' (instead of 'Appen Butler Hill Incorporated'), the word 'Inc' is transcribed.

5.4.3 Punctuation

Punctuation marks are not used in transcription unless they are an essential part of the word. For example:

"can't"

5.4.4 Acronyms

Acronyms are transcribed as words if spoken as words, and as letters if spoken as letters. When transcribing sequences of letters an underscore is inserted between each letter. For example:

NASA I_B_M

5.4.5 Numbers

Numbers are transcribed as full words. For example:

16 -> sixteen
112 -> one hundred and twelve

5.4.6 Phonetic Spelling

The / / tag is used for letters that are pronounced as the sound, rather than as the word. For example, when a person means to convey the letter B but they say the sound 'b' instead of the word 'bee'. In this case we transcribe it as /B/.

6 Pronunciation Lexicon Specifications

The specifications below detail the approach and content of the pronunciation lexicon.

6.1 Objective for the Pronunciation Lexicon

The main function of the lexicon is to provide a pronunciation dictionary for the corresponding language corpus. The technology developers can associate with each utterance the most probable phonetic transcription in the absence of a detailed, manually provided phonetic transcription of each utterance. While syntactic and semantic information could be of great benefit in identifying the correct words and corresponding pronunciations for each word occurring in the database, it is more efficient to list all word forms which occur in the corpus with their corresponding canonical phonemic transcription. The intention is therefore to list all fully inflected forms encountered, and not the stem and morphological components. It is therefore up to the user to ensure that the correct word forms are identified and applied in using the material.

Although the general approach to pronunciation lexica is to produce a phonemic pronunciation, allophonic variations are permitted in the lexicon and would be documented in the phone set.

6.2 Handling of Orthographic and Pronunciation Variants

There are 5 special cases that need to be considered when preparing the lexicon:

homonyms	2 words with the same orthography and phonological form but different syntax and/or meaning e.g. “mate” (a friend, a chess position, to join together)
homographs	2 different words with the same orthography but different phonology e.g. “read” /ri:d/ or /rEd/
homophones	2 different words with different orthography but identical phonology e.g. “bred” and “bread”
heterophones	2 or more phonological forms for the same word (multiple pronunciations) e.g. “either”- /aI.D@/ or /i:.D@/
heterographs	2 or more orthographic forms for the same word e.g. “recognise” and “recognize”

As Appen was primarily concerned with providing the full set of possible phonetic transcriptions for each word form, and is not encoding syntax or semantic information, they minimally supply entries for all word forms encountered, and treat each of the cases above as follows:

1. homonyms are not distinguished (indeed they cannot be in our tables);
2. homophones are easily distinguished and should be listed as separate lexical items containing the same pronunciation;
3. homographs should be listed if present in the corpus and presented as pronunciation variants;
4. heterophones should ideally be listed wherever they are known and presented as pronunciation variants;
5. heterographs should not occur as standard dictionaries and spellings.

With the exception of homographs and heterophones, there is a one-to-one mapping between each word and its canonical phonemic transcription.

6.3 Structure and Conventions

The lexicon file is in UTF-8 delimited format with a TAB character as field delimiter. The lexicon is **an alphabetically ordered** list of distinct lexical items (essentially words in our case) that occur in the corpus with the corresponding pronunciation information.

If the words are in non-Latin script the locale used for sorting these UTF-8 files is by default LANG=c, unless there is a language-specific locale that is more appropriate. The lexicon entries must be identical to the transcription words.

The pronunciation information should be in the appropriate SAMPA format.

6.4 Syllabification and Stress Coding

Words are syllabified according to principles relevant to the particular language. Stress information and word/morphological boundary information are each individually optional but, if included, would follow the conventions tabled below. The stress marks " and % are written immediately before the syllable. All these conventions are space-delimited and have been designed to permit the removal of information that is not required, or the selection of useful subsets of the table.

Word in compounds	#
Syllable	. (period)
primary stress	“ (double quote)
secondary stress	% (percentage mark)

Table 6-1 Conventions for marking boundaries and stress

6.5 Apostrophes and Hyphens

Words in the lexicon are only split at spaces, not at hyphens or apostrophes.

6.6 Lexicon Format

The lexicon is a multi-column text file. An example of a small English lexicon file is in the table below (“\t” represents the TAB character):

WORD		PHONEMIC TRANSCRIPTIONS				
chin	\t	" tS I n				
cut	\t	k V t				
either	\t	" aI . D @	\t	" i: . D @		
heard	\t	"h e r d	\t	"h 3: d	\t	"3: d
mock	\t	" m Q k				
pin	\t	" p I n				
read	\t	" r i: d	\t	" r E d		
red	\t	" r E d				
thin	\t	" T I n				

Table 6-2 Example of English lexicon file

For non-Latin script languages (like Chinese, Japanese, Russian, etc.) the lexicon contains both original and Romanized characters in parallel. For tonal languages, tonal markers occur in the phonemic transcription and are separated by an underscore. The representation of these tonal markers is specified in the LSP document.

WORD		ROMANIZED WORD		PHONEMIC TRANSCRIPTION
仲谋	\t	zhong4mou2	\t	ts` w 7 N _F . m 7w _R
红鸟	\t	hong2niao3	\t	x w 7 N _R . n j aw _FR

Table 6-3 Example of Chinese (Mandarin) lexicon file

For Arabic scripted languages or other languages where the vowels are generally not coded in the words (like Hebrew, Farsi, Urdu, Pashto, etc.) the lexicons consist of the unvowelized word, an optional vowelized word, the Romanized form, then the phonemic transcription, as in Table 6-4 below.

WORD		ROMANIZED WORD		PHONEMIC TRANSCRIPTION
الخباز	\t	Alxab~aAzo	\t	? a l . x a b . " b a: z
الراية	\t	AlraAyap	\t	? a r . " r a: . j a
سكنة	\t	sakonap	\t	" s a k . n a
سكنة	\t	sakonap	\t	" s a k . n a t

Table 6-4 Example of Arabic (MSA) lexicon file

7 Data Delivery Format

As stated in the BAA, we will provide training and development data sets for all of the languages explicitly addressed in the program. Separate data sets will be used for the evaluations. The ***training and development data*** for each language will be disseminated in a system ***build pack*** for that language. CRC checksum has been run to validate directories and files. The pack will contain a description of the language, including limited information on the dialectal variation in the collection, and the phoneme set for the language. Additionally, it will contain the audio recordings and transcription of some portion of that audio using a normalized orthography.

- Training data will contain both scripted/prompted short spoken utterances (e.g., short responses to questions, phone numbers, dates) to ensure coverage of the language's phoneme inventory, and also conversational speech between pairs of people who already know each

other, conversing on any topic(s) they choose with the percentage of transcription depending on the Program Period. A pronunciation lexicon of words appearing in transcribed training data will be provided in extended SAMPA.

- Ten hours of transcribed development data will be provided to measure interim progress. These data must not be used for system training, but can be used for optimization of system parameters.

7.1 Directory Structure

7.1.1 Top level directory

Language build packs are delivered to the principal investigators (PIs) on USB flash drives. The top level directory structure defined in Table 7-1 is used for all language deliveries of the program.

Directory	Description
<language_pack/>	Directory named with specific Babel language code

Table 7-1 Top Level Directory Structure

Language pack directories have a unique name corresponding to the identity of the language. The top level directory is named using the following label convention:

BABEL_<PERIOD>_<LANGUAGE CODE>

Where
<PERIOD> is a two character representation of the year of the Babel program. For example "BP" for Base Period or "O1", "O2", "O3" for the follow on option periods of the programs;

<LANGUAGE CODE> is a three digit code that is used to represent the language in the package.

7.1.2 Language_pack directory

Language packages contain two types of speech data that have been collected by Appen Butler Hill. Conversational data involves two speakers speaking spontaneously on any subject they wish without a script. Scripted audio contains short spoken utterances, e.g., short responses to questions, phone numbers, dates, as well as phrases to ensure coverage of the language's phoneme inventory. Files from each format will be placed into separate directories as indicated in Table 7-2.

Directory	Description
<language_pack/>	Directory named with specific Babel language code
conversational/	Directory containing conversation speech data.
scripted/	Directory containing scripted speech data.

Table 7-2 Language Pack Directory Structure

7.1.3 Speech data directories

The folder for each type of speech data (conversational and scripted) is partitioned into five subfolders in order to organize the training data (training), development data (dev), evaluation data (eval), sub-train data, and reference materials (see Table 7-3). Four of the subfolders contain data for that specific partition type (described in Section 7.1.4). The eval partition folder will be empty for system build deliveries, but can be populated once the evaluation data is released. The sub-train folder consists of a

10 hour subset of training data. Note that since the audio files exist in training, they are not repeated in the sub-train directory. Section 7.2.3 provides details on the reference material subfolder.

Directory	Description
scripted/ **or** **conversational/**	Directory named with specific Babel language code
training/	Subdirectory containing the training data set for model development
dev/	Subdirectory containing the development partition data set for performance analysis
eval/	Subdirectory structure set up for storing the test data for evaluation (conversational only)
sub-train/	Subdirectory containing a 10 hour subset of training data
reference_materials/	Support documentation for the project

Table 7-3 Data Format Directory Structure

7.1.4 Partition directory

Each partition folder contains two or three directories, one for each of the different types of files that support the language pack. These directories are defined in Table 7-4 and are described in more detail in Section 7.2. Depending on whether or not the language requires Romanization, transcript_roman may be omitted.

Directory	Description
partition_type/	training/dev/eval directory
audio/	Subdirectory containing audio files
transcription/	Subdirectory containing UTF-8 versions of the transcription files (scripted/conversational; empty for the eval partition)
transcript_roman/	Subdirectory containing Romanized versions of the transcription files (scripted/conversational ; empty for eval partition)

Table 7-4 Partition Directory Structure

7.2 Data File Formats

7.2.1 Audio encoding and formats

Audio is encoded as sequences of 8-bit 8-KHz A-law or mu-law speech samples. Each channel within a conversational call is stored within a separate file. Note that channel files from the same session may not necessarily be time synchronized. Files are distributed with a SPHERE audio file standard header that has been defined by the National Institute of Science and Technology (NIST) and inserted using a standardized tool suite[1].

7.2.2 Transcription data

Transcription files are text files where each line ends with a <CR><LF> sequence.
The transcription is done following segmentation of the signal files into utterances. Each utterance is transcribed on a new line beginning with a time-stamp that indicates the beginning of the utterance. The time-stamp appears in square brackets. For example, [000.000] is the initial time stamp in the file.

[1] Language Technology Tools - http://www.nist.gov/itl/iad/mig/tools.cfm

The transcription file contains only the timestamps, transcriptions, and tags. See Section 5 for more information on transcription conventions, including a list of tags.

Transcriptions are provided for some portion of the audio in the train directory, and all of the audio in the dev directory. Associated transcription files are formatted as UTF-8 text files. An example of the format is shown in Figure 7.1. This is provided for example purposes; the portions marked <conversation> would be replaced with the actual transcription text in a delivery.

```
[0.000]
<no-speech>
[3.010]
<no-speech>
[10.640]
<no-speech>
[17.640]
#silence <conversation> #breath
[25.830]
<Conversation><no-speech>
[33.890]
<conversation><no-speech><conversation><no-speech>
[40.290]
<conversation>
[46.020]
<conversation><no-speech>
…
…
…
[600.260]
```

Figure 7.1 Example Transcription File

7.2.3 Reference material

The reference material folder consists of the lexicon and the metadata information.

7.2.3.1 Lexicon data

A file named “lexicon.txt” includes the words from the transcription files in training and dev and their corresponding pronunciations. Please note that a lexicon for words unique to dev is not provided. Another file “lexicon.sub-train.txt” includes the words from the transcription files in sub-train and their corresponding pronunciations.

7.2.3.2 Metadata

A file named “demographics.tsv” includes metadata information corresponding to the audio files in training and dev. Another file “demographics.sub-train.tsv” includes metadata information corresponding to the audio files in sub-train. There are several fields of metadata that may be useful in support of the algorithm development tasks. This metadata will be provided for both conversational and scripted audio files. A tab separated file, one for conversational audio and one for scripted audio named “demographics.tsv” will be provided containing the following pieces of information:

- File Name (outputFn): See Section 7.2.4.
- Session ID (sessID): a code associated with a recording session during collection; it is a five-digit number ranging from 00000 up to 99999.
- Date (date): recording date
- Time (time): recording time
- Speaker Code (spkrCd): unique speaker code (this code is usually the same as the sessID except in some rare cases).
- Line Type (lineType): inLine or outLine for conversational, inline only for scripted
- Dialect (dialect): speaker accent, i.e., the regional/dialectical colouring factor

- Gender (gen): speaker sex ('M'ale or 'F'emale; void if unknown)
- Environment (envType): Calling environment; see Table 7-5.
- Age (age): speaker age (precise or class mid-point ; void if unknown)
- Network (network): Telephone network
- Telephone model: telephone hand set model

Figures 7.2 and 7.3 show an example of the metadata for conversational audio and scripted audio, respectively.

outputFn	sessID	date	time	spkrCode	lineType	dialect	gen	envType	age	network	phoneModel
BABEL_BP_101_48053_20111020_130943_outLine.sph	48053	20111020	130943	48053	outLine	Central_Guangdong	M	HOME_OFFICE_MOBILE	53	CMCC	Etonmobile
BABEL_BP_101_55786_20111023_175604_inLine.sph	55786	20111023	175604	55786	inLine	Southern_Pearl_River_Delta	F	STREET	23	CMCC	Nokia
BABEL_BP_101_84943_20111020_144955_inLine.sph	84943	20111020	144955	84943	inLine	Guangxi_and_Western_Guangdong	F	HOME_OFFICE_LANDLINE	30	China_Telecom	Panasonic
BABEL_BP_101_16313_20111022_221750_outLine.sph	16313	20111022	221750	16313	outLine	Southern_Pearl_River_Delta	M	HOME_OFFICE_MOBILE	29	CMCC	Samsung

Figure 7.2 Example of Metadata file for Conversational Audio

outputFn	sessID	date	time	spkrCode	lineType	dialect	region	gen	envType	age	network	phoneModel
BABEL_BP_101_90506_20111026_180107_C5_scripted.sph	90506	20111026	180107	90506	inLine	Guangxi_and_Western_Guangdong	Guangzhou,China	F	CAR_KIT	46	CMCC	Philips
BABEL_BP_101_42615_20111018_171511_S8_scripted.sph	42615	20111018	171511	42615	inLine	Central_Guangdong	Guangzhou,China	F	HOME_OFFICE_MOBILE	22	China_Unicom	Sony_Ericsson
BABEL_BP_101_81308_20111021_141809_S1_scripted.sph	81308	20111021	141809	81308	inLine	Central_Guangdong	Guangzhou,China	M	PUBLIC_PLACE	59	China_Telecom	Nokia
BABEL_BP_101_65743_20111019_155019_L1_scripted.sph	65743	20111019	155019	65743	inLine	Southern_Pearl_River_Delta	Dongguan,China	F	HOME_OFFICE_LANDLINE	50	China_Telecom	BBK

Figure 7.3 Example of Metadata file for Scripted Audio

EnvType	Description
HOME_OFFICE	calls originated from a quite location, such as home, office, etc; they can be affected by background homely noises such as TV, children, background talking, telephone rings, typewriting noises
PUBLIC_PLACE	calls generated by using telephones located in public places, with high background talking, such as: station, airport, bar, kiosk
STREET	calls generated by using telephones from sites with high background traffic emission noise
VEHICLE	calls generated by using cellular phones from inside moving vehicles such as: car, lorry, bus, coach, tram and train
CAR_KIT	calls coming from a passenger in moving car using an installed car kit with a far-field microphone (mic placed >20 cm. from speaker's mouth)
OTHER	surprise environment type different from all the above

Table 7-5 Environment types (excluding Unknown types that may appear in evaluation)

7.2.4 File naming convention

Each language package will use file naming conventions for file names. The following naming convention is used for conversational files with the variables defined in Table 7-6.

ConvFilename=BABEL_<PERIOD>_<LANGUAGE_CODE>_<SESSION_ID>*_<YYYYMMDD>_<HHMMSS>_<LINE>.<FILE TYPE>*

An example of a conversational audio file with this convention is:

BABEL_BP_101_ 97016_20110917_124247_inLine.sph

File Label	Description
<PERIOD>	A two character representation of the period of the Babel program
<LANGUAGE_CODE>	3 digit code which represents the language
<SESSION_ID>	A 5 digit code that Appen Butler Hill uses for recording session identification
<YYYYMMDD>	Recording date (e.g. 20110910 for 10 Sep 2011)
<HHMMSS>	Recording time (e.g. 112351 for 11:23:51 AM, 232351 for 11:23:51 PM)
<LINE>	The channel recorded, where inLine is the caller outLine is the call receiver
<FILE TYPE>	sph: audio files with sphere header txt: text files containing the transcription

Table 7-6 Variables for Conversational File Naming Convention

The following naming convention is used for scripted files with the variables defined in Table 7-7.

ScriptFilename=BABEL_<PERIOD>_<LANGUAGE_CODE>_<SESSION_ID>*_<YYYYMMDD>_<HHMMSS>_<CORPUS_CODE><LINE>.<FILE TYPE>*

The corpus code is the concatenation of the corpus identifier and the item identifier as described in Table 2-1. An example of a scripted audio file with this convention is:

BABEL_BP_101_ 97016_20110917_124247_A1_scripted.sph

File Label	Description
<PERIOD>	A two character representation of the period of the Babel program
<LANGUAGE_CODE>	3 digit code which represents the language
<SESSION_ID>	A 5 digit code that Appen Butler Hill uses for recording session identification
<YYYYMMDD>	Recording date (e.g. 20110917 for 17 Sep 2011)
<HHMMSS>	Recording time (e.g. 112351 for 11:23:51 AM, 232351 for 11:23:51 PM)
CORPUS_CODE	Two character representation of the Corpus Code of the scripted material (see Table 2-1)
<LINE>	scripted - has been placed in the file name to indicate it comes from a scripted recording
<FILE TYPE>	sph: audio files with sphere header txt: text files containing the transcription

Table 7-7 Variables for Scripted File Naming Convention

7.3 Language Specific Peculiarities

Language-specific details are described in a Language Specific Peculiarities (LSP) document and are also provided with the delivery. They will be placed on the SharePoint site. The topics appearing in these documents are:

- Dialect Regions
- Deviation from Native-Speaker Principle
- Special Handling of Spelling
- Description of Character Set Used for Orthographic Transcription
- Description of Romanization Scheme
- Description of Method for Word Boundary Detection
- Table Containing All Phonemes in the Stipulated Notation
- Complete List of All Rare Phonemes
- Other Language Specific Items
- References

8 Bibliography

[1] Richard Winski, "Definition of corpus, scripts and standards for Fixed Networks", SpeechDat project, doc. ref. LE2-4001-SD1.1.1, 22 July 1996

[2] J. G. van Velden, "Specification of Speech Data Collection over Mobile Telephone Networks", SpeechDat project, doc. ref. LE2-4001-SD1.1.2/1.2.2, 14 October 1996

[3] F. Senia et all, "Environmental and speaker specific coverage for Fixed Networks", SpeechDat project, doc. ref. LE2-4001-SD1.2.1, 1997

[4] F. Senia, Jeroen G. van Velden, "Specification of orthographic transcription and lexicon conventions", SpeechDat project, doc. ref. LE2-4001-SD1.3.2, 16 January 1997

[5] F. Senia, "Specification of speech database interchange format", SpeechDat project, doc. ref. LE2-4001-SD1.3.1, 28 February 1997

[6] G. Chollet, F. T. Johansen, B. Lindberg, F. Senia, "Test Set Definition and Specification", SpeechDat project, doc. ref. LE2-4001-SD1.3.4, December 1997

[7] Dutch Polyphone spec and transcription guidelines

[8] SIL Three-letter Codes for Identifying Languages: http://www.ethnologue.com/codes

[9] Tags for the Identification of Languages: http://www.ietf.org/rfc/rfc3066.txt

[10] Codes for the Names of the Languages: http://www.loc.gov/standards/iso639-2/englangn.html

[11] Internet Assigned Numbers Authority: http://www.iana.org/assignments/character-sets

[12] Asunción Moreno, Francesco Senia, "LILA project specifications", V2.3, 20 February 2006