Speech Technologies: Language Variation

Language variation in speech technologies

Simon King

August 14, 2004

Abstract

Spoken language technologies, such as automatic speech recognition, must often deal with speech from many speakers and in a wide variety of situations. Variation in speech usually creates serious difficulties for such systems; this chapter looks at the many sources of variation in speech, within a single speaker, across speakers and languages, and in external factors.

1 INTRODUCTION

This article deals only with the sources and types of variation in speech that affect speech technology systems such as ASR and TTS. A companion article, *Handling variation in speech and language processing*, looks at how such systems attempt to deal with this variation. Since this article is therefore largely a survey of well-understood linguistic phenomena, almost all the references will be to separate articles in this encyclopedia. These articles give suggested reading beyond that given here: *Handling variation in speech and language processing*; *Speech Technologies, Overview*; *Natural Language Processing, Overview*; *Speech Recognition, Statistical Methods*.

1.1 Speech recognition

Automatic speech recognition, **ASR**, (*see* Speech Recognition, Statistical Methods) usually means the conversion of speech waveforms into sequences of words (i.e. text), and does not usually imply any further syntactic (*see* Parsing, statistical methods; Parsing, Symbolic) or semantic (*see* Natural Language Understanding, Overview) processing of this text, such as might be necessary in a spoken dialogue system (*see* Computational Language Systems, Architectures).

Most modern ASR systems have a common architecture. The speech waveform is invariably preprocessed into a sequence of vectors (called feature vectors), each representing the short-term spectral envelope of a section of waveform. Then a hierarchy of statistical models, usually consisting of a model of word sequences – the language model, usually an **n-gram model** –, a word-word sequence. Statistical models outperform alternatives, such as rule-based or example-based systems. In the case of HMMs, this is because they are able to learn, from data, not only the average value of the feature vectors associated with a particular unit of speech, but also the *variation* about this average. HMMs are trained on large quantities (several 100 hours) of transcribed speech data and N-gram models are trained on very large quantities of text (e.g. more than 100 million words). to-phone model – the pronunciation **lexicon** (*see* Computational Lexicons and Dictionaries) – and models of phone-sized units of speech – usually **hidden Markov models, HMMs**, (Huang, Acero & Hon, 2001) – is used to determine the most probable word sequence. Statistical models outperform alternatives, such as rule-based or example-based systems. In the case of HMMs, this is because they are able to learn, from data, not only the average value of the feature vectors associated with a particular unit of speech, but also the *variation* about this average. HMMs are trained on large quantities (several 100 hours) of transcribed speech data and N-gram models are trained on very large quantities of text (e.g. more than 100 million words).

Given a good enough model, and sufficient quantities of appropriate training data, statistical models are able to **generalize**: they can recognize new speech data with word sequences never seen during training.

However, when presented with speech which differs too greatly in some respect from the training data, the accuracy can drop dramatically. Understanding the types, sources and acoustic consequences of variation in speech is therefore important to the developers of ASR systems.

In the early development of ASR systems (*see* Speech Recognition, Automatic, History of), the focus was on transcription of so-called “read text”, since the primary application of the technology was thought to be dictation machines. More recently, interest has shifted to processing spontaneous speech as found in dialogue (whether between humans or human and machine). Challenging problems currently under investigation include: transcription of spontaneous telephone conversations, “rich transcription” of speech, transcription and annotation of multi-party meetings and searching of speech databases (*see* Spoken Document Retrieval, Automatic).

1.2 Speech synthesis

The conversion of plain text (such as the paragraph you are reading now) to speech waveforms is known as **text-to-speech, TTS,** (*see* Speech Synthesis). When more information about the text is available (such as the correct syntactic structure, or semantic information such as what is given and what is new information in a particular utterance), the task is then sometimes called **concept-to-speech**. This extra information will be available if the text was automatically generated (*see* Natural language generation, overview). Clearly, TTS is the harder problem since any syntactic, semantic or pragmatic information required by the system must be inferred from the text (*see* Parsing, Symbolic).

This is usually achieved with a combination of rules and statistical models (Jurafsky & Martin, 2000) Most TTS system construct the synthetic speech waveform using fragments of speech taken from a database of natural recorded speech (from a single speaker); these systems are known as **concatenative**.

The variability and ambiguity present in text can present problems for TTS systems in two ways. Firstly, a single sentence taken out of context will often appear to have more than one "correct" reading; this ambiguity will need contextual information to resolve, which may be available in the preceding sentences, or may require knowledge beyond that contained in the text (world knowledge). Typical TTS systems have very little world knowledge. Secondly, the rules and statistical models used in typical TTS systems are usually optimized or trained on some data set; however large it is, this data set cannot cover all possible words and phrases that will be encountered when running the system. So, just like for ASR, a TTS system must be able to **generalize** from limited training data.

1.3 Other speech technologies

As well as recognizing it and synthesizing it, there are other interesting and useful things we can do with speech.

When transmitting speech down a long-distance telephone cable, or through the airwaves to and from mobile phones, it is usual to **code** the speech. This coding achieves both substantial compression (a lower data rate than the original waveform) and robustness to errors introduced by the channel. The coding is usually specifically designed for speech; for example, it might take advantage of a source-filter model and code the source and filter components separately. Since a model of speech is being used (however simplistic), there are potential problems with variation, particularly across speakers.

In **voice transformation** (also known as voice morphing), the speech of one speaker is manipulated to sound like another speaker (*see* Voice Modification, Synthetic). This has applications including speech synthesis and film dubbing. Clearly, understanding the factors that make one person's speech differ from another's is crucial when performing voice transformation.

Both ASR and TTS are required for a **spoken language dialogue system**. In such a system, some form of speech understanding must be implemented (*see* Speech Understanding, Automatic) . If this involves parsing the input, then the differences between spoken and written language must be taken into account, since most parsers are developed for written language only.

Differences between speakers are not always a problem. In the task of speaker recognition, these differences are used to identify, or verify a claimed identity of, the speaker (*see* Speaker Recognition and Verification, Automatic).

2 VARIATIONS IN THE SPEECH OF A SINGLE SPEAKER

No speech sound is ever produced in precisely the same way twice, even by a single speaker; the number of dimensions along which we can describe, and perhaps quantify, this variation include:

Speech production processes We can divide variations in the speech production process into two distinct categories. The first consists of systematic variations due to phonetic context, including allophonic variation and coarticulation. The second category consists of random variations.

Allophonic variation is the systematic variation in the acoustic realization of a phoneme depending on the phonetic context. Examples include: a stop may be released in some situations but not in others (such as before another stop); /l/ in English may be realized differently in syllable-initial vs. syllable final positions ("light" vs. "dark"). Clearly, such systematic variation should be accounted for in ASR (e.g. by using a different HMM for each type of /l/) and in TTS (by synthesizing the appropriate /l/ in each context).

Coarticulation (*see* Articulatory Phonetics) is the variation in underlying articulator movements depending on context, both left and right. For example:

[h æ n d]	+	[b æ g]	→	[h æ m b æ g]
hand		bag		"hambag"

illustrates the process of **assimilation** (*see* Assimilation) where the [n] has assimilated the place of articulation of the following bilabial stop [b], thus changing an alveolar [n] to a bilabial [m]. The [d] has also been deleted. These processes must be realized in TTS and accounted for in ASR. Cross-word effects, such as the example above, cannot be handled in the lexicon.

The second category of variation is that of random changes. These arise because the motor control system of the articulators is imperfect (we cannot exactly repeat an articulatory gesture). Human speech perception has no problem with the acoustic consequences of small random changes since they do not cause categorical changes (*see* Speech perception). Because these changes are not predictable, in ASR they must be left to the low-level statistical component (usually a mixture of Gaussian distributions) to deal with. In TTS, including these variations in the synthetic speech may improve naturalness.

Speaking rate At higher speaking rates, the processes mentioned above have an increasing effect on the speech produced. Co-articulation becomes more pronounced and the frequencies of deletions, disfluencies, mispronunciations all increase.

3 DIFFERENCES BETWEEN INDIVIDUAL SPEAKERS

The speech of two speakers of the same language will exhibit systematic differences, not attributable to external factors.

It is common in speech technology to make use of the source-filter model (*see* Acoustic Phonetics; Speech Production) which divides the speech production process into two components: the source (e.g. the vocal folds for voicing) and the filter (the vocal tract). This model is somewhat simplistic, but is widely used in speech technologies, and is useful in understanding cross-speaker variation.

Differences in vocal tract geometry, particularly in length, cause shifts in the formant space; a shorter vocal tract (e.g. in female speakers compared to male speakers of the same age, or children compared to adults) produces higher formant frequencies. Since the acoustic features used in ASR systems represent the spectral envelope (an approximation to the vocal tract frequency response), the features extracted for different speakers will vary systematically.

The range of F_0 (the rate of vibration of the vocal folds, whose perceptual correlate is pitch) varies across speakers too. This creates challenges for some ASR systems and all TTS systems. In ASR of non-tone languages, F_0 is almost always ignored; however, in tone languages, where F_0 carries segmental information it is necessary to model F_0; this raises the problem of normalizing across speakers: one Cantonese speaker's Low Rising tone might have a very similar F_0 contour (in both absolute value and in shape) to another speaker's Mid Rising tone (*see* Tone, Phonology of).

The vocal folds (*see* Speech Anatomy; State of the Glottis) are also largely responsible for voice quality (*see* Voice Quality). In model-based TTS (*see* Speech Synthesis), this means a detailed model of the vocal fold behavior is necessary to capture the individual qualities of any real speaker. In voice transformation (*see* Voice Modification, Synthetic), conversion of this aspect of speech is critical,

4 VARIATION IN SITUATION

Speakers change many aspects of their speech, both consciously and subconsciously, in response to situational factors.

4.1 Planned vs. spontaneous speech

There are important differences between read text (e.g. that of a newscaster), careful planned speech (e.g. from an experienced public speaker without notes) and spontaneous speech (e.g. a conversation).

4.2 Effect of the listener on the speaker

Studies of human-human dialogue, such as those of the Maptask corpus (Anderson et al, 1991), have been used to discover the effects of speaker familiarity and eye contact between speakers. Other factors that affect speakers include their relationship to the listener (their child, their boss,...). Interestingly, when speakers know the listener is a machine, they are more likely to adapt their speech (to become more careful) than if they believe they are talking to a human. This has implications for spoken dialogue system designers: if the system is so good that users believe it is a human (*see* Turing Test), their speech is more likely to be problematic for the ASR component.

4.3 Disfluencies and related phenomena

Word fragments, repeated word or phrases and repairs cause problems for ASR systems. Since ASR systems almost always see speech as a string of words to be transcribed, and use a simple N-gram model of language, the insertion of non-words or word fragments into the word string cause problems. Should they be transcribed? Can they be treated as real words? How do they affect the language model? *Handling variation in speech and language processing* considers these questions. Hesitations and filled pauses are not always speech errors; they may be used by speakers, particularly in dialogue, to perform functions such as holding the current turn.

4.4 Dialogue vs. monologue

In dialogue speech (or speech between two or more speakers), there are significantly more disfluencies than in monologue. Speakers overlap their speech significantly; in a small meeting situation, we may find as many as half of all utterances overlap. Even when recording the speech with close-talking head-mounted microphones, it is very difficult to obtain separate acoustic signals for each speaker. This presents a major challenge to speech recognition systems.

5 VARIATIONS WITHIN A SINGLE LANGUAGE

All languages with large numbers of speakers contain sub-groups: accents (*see* Accent) and dialects (*see* Dialect Atlases).

Variations between different accents include systematic and idiosyncratic changes in word pronunciations. For example, accents of English can be divided into rhotic and non-rhotic varieties. Speakers with rhotic accents (e.g. Scottish English, most accents of North-American English) pronounce the [r] at the end of words such as “nicer”, whereas non-rhotic speakers (e.g. of Southern

British English) do not. In concatenative TTS systems (*see* Speech Synthesis) particularly, it is important for the pronunciation dictionary to closely match the speaker whose speech is being used.

Different dialects of a language can exhibit all the differences between different accents, plus differences in vocabulary (*see* Standard and dialect vocabulary) and syntax. Since speech technology systems tend to be constructed for "standard" accents/dialects, "non-standard" speakers are more likely to have difficulty using these systems. The same applies for non-native speakers.

6 VARIATIONS BETWEEN LANGUAGES

Until recently, only a small number of languages had been used in the primary research and development of speech technology; of these, English, particularly North-American English, dominated (and still does). When attempts are made to "port" existing systems to other languages, a number of cross-language differences must be dealt with. Some differences are minor and the system can ported simply by replacing some of the linguistic resources (e.g. the pronunciations dictionary) or providing new training data. Other differences are more significant, and the system must be changed in a more fundamental way. Systems that must deal with more than one language may require to automatically detect which language is being spoken (*see* Language Identification, Automatic).

6.1 Across closely-related languages

Phoneme inventory Speech technology systems typically use relatively small phoneme inventories compared to the full set of IPA (*see* International Phonetic Association) phonemes. First, almost all systems are designed to deal only with a single language at a time, so only a subset of the IPA phonemes are required. Second, small inventories are inherently preferred. In ASR, this is because a statistical model must be built for every phoneme in every left and right context. If there are 40 phonemes and the context consists of only one phoneme to the left and one to the right, this means around $40^3 = 64000$ models; with 50 phonemes this rises to 125000. In TTS, the typical fragment of speech used by concatenative systems is the **diphone**, consisting of the second half of one phone plus the first half of the following phone. Hence, the number of diphones increases as the square of the number of phonemes.

Morphology Both ASR and TTS require pronunciations for words. In TTS, only the orthographic form is available as input; any extra information required to predict pronunciation (such as part-of-speech (POS) tags (*see* Part-of-speech tagging)) must be inferred by the system. This process of **letter-to-sound (LTS)** conversion (*see* LTS) must be automatic since new words will be encountered at run-time; for large-vocabulary ASR, (semi-)automatic methods are also desirable. For languages with productive morphology, automatic morphological decomposition (*see* Finite-state Methods, Morphology, Computational) can be used to assist LTS, since the pronunciations of roots and affixes can be shared across many words and combined to produce pronunciations for complete words. Languages for which morphology is commonly used in LTS include Spanish, German and Arabic; English, for example, is too irregular. Since systems for English rarely use a morphological component, this must be added when porting to most other languages. (*see* Dictionaries and inflectional morphology, Morphology and language processing).

Lexicon and vocabulary size Languages with a highly productive morphology can have very large vocabulary sizes. The standard approach used for English ASR and TTS, which is to treat words as atoms and list all (orthographically) distinct words in the lexicon, is not appropriate for such languages. In ASR for these languages, failing to make use of morphology will mean a much higher "out of vocabulary" rate: words appearing in speech to be recognized that are not listed in the pronunciation dictionary of the system.

Languages with writing systems that omit some vowels, or do not mark word boundaries, create additional problems for speech technologies. New modules must be added to the system to predict these missing pieces of information before words can be looked up in the lexicon.

Syntax In ASR, there is rarely a sophisticated model of syntactic structure. Language is modelled a string of words with no deep structure, so therefore differences in syntactic structure cause no real problems, although the power of simple n-gram models will vary across languages; for example, n-gram models cannot capture long-range dependencies.

In TTS, however, some syntactic analysis is usually performed. This may be very shallow, such as part-of-speech tagging (*see* Symbolic computation linguistics, Overview; Parsing, Symbolic) or somewhat deeper – to discover phrase boundaries, for example. In English, for example, the relationship between traditional syntactic constituents (e.g. noun phrases, verb phrases) and prosodic structure (*see* Prosodic Aspects of Speech and Language) is via semantics (*see* Semantics of prosody), so even a deep syntactic analysis is not sufficient. One theory of syntax, Combinatorial Categorical Grammar (CCG) (*see* Combinatory Categorial Grammar), is able to generate constituents which better match the information structure, and therefore the prosodic structure, of utterances.

6.2 Across distantly-related languages

Moving on to more distantly related languages, further problems are encountered.

Differences in acoustic features Tone languages use F_0 to make segmental distinctions. A speech recognition system for a non-tone language will generally not use F_0 information, so when ported to a tone language, the so-called “front end” of the system, that extracts salient acoustic features from the waveform, will need modifying.

Languages without existing linguistic resources For some languages there are few or no existing resources, such as large amounts of text data in electronic form or pronunciation dictionaries. These resources must be created, and this is one of the largest barriers to use of speech technologies for these languages. Projects such as the Local language speech technology initiative (Tucker, 2004) are addressing this problem.

Differences between spoken and written language Even widely spoken dialects of major languages may lack linguistic resources. For example, Levantine Arabic (ethnologue name: Arabic, Levantine Bedawi spoken) is a spoken dialect of Arabic. Since writing in dialect is stigmatized, written Arabic is almost exclusively Modern Standard Arabic, which is significantly different (being a modernized version of Classical Arabic).

7 NON-LINGUISTIC EXTERNAL FACTORS

A number of factors can modify speech signals on the way from speaker to speech technology system.

7.1 Transmission factors

Telephone speech Telephone networks generally limit the bandwidth of the speech signal they carry (typically limiting the range of frequencies present in the signal to 300-3500Hz). This reduces the amount of information available to ASR systems. For many male speakers, this also means that the first harmonic (sometimes called the fundamental) of F_0 is lost. Human listeners are able to infer F_0 from the spacing of the remaining harmonics.

Digital channels (which includes most mobile and long distance channels), impose degradation on the speech signal due to the coding and decoding of the signal (Gold and Morgan, 1999). Mobile telephones are frequently used in noisy (e.g. in-car) and reverberent (e.g. subway station) environments.

Microphone Different microphones can have surprisingly large effects on the speech signal, to the extent that ASR accuracy can be severely reduced when the microphone used in testing is different to that used to collect the training data.

7.2 Environmental factors

Even before the speech reaches the microphone, it can be corrupted.

Acoustic (background) noise A particular problem of far-field microphones (e.g. desk-mounted vs. head-mounted) is the higher level of background noise, relative to the speech signal.

Reverberent environments When multiple paths between speaker and microphone are present (e.g. because of reflections from walls), the signal received by the microphone contains not only the direct-path signal, but also weaker, delayed versions. This "blurs" the information across time and can cause significant degradation in speech quality. In making recordings for concatenative TTS systems, great care is taken to avoid recording these reflected signals (by using a recording studio or anechoic chamber). Far-field microphones also pick up a higher level of these reflected signals than close-talking microphones, relative to the direct-path speech signal.

Bibliography

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R. (1991). 'The HCRC map task corpus.' *Language and Speech, 34*, 351–366.

Huang, X., Acero, A., Hon, H-W (2001). *Spoken language processing: a guide to theory, algorithm and system development*. Prentice Hall.

Gold, B. & Morgan, N. (1999). *Speech and Audio Signal Processing*. Wiley.

Jurafsky, D. & Martin, J., (2000). *Speech and language processing*. Prentice Hall.

Tucker, R. (2004). `http://www.llsti.org`