nexidia

FACT SHEET

Nexidia's Phonetic Based Speech Analytics: A Robust Solution to Language Variation

INTRODUCTION

Dialectal variation poses a number of challenges for speech analytics. In many regions of the world, there is no clear distinction between a language and a dialect. In addition, people may speak different dialects/languages depending on the situation and the addressee, or they may even mix different dialects. In these situations, the phonetic based algorithms developed by Nexidia provide a much more robust way to perform speech analytics than automatic transcription ("speech-to-text") approaches.

DIALECT CONTINUA

When most people think of languages, they think of the standardized forms of languages spoken by specific nations. Thus, in France they speak French, and in Italy they speak Italian. In reality, however, it is much more difficult to tell where one language begins and another ends. As you cross from one country into another, it is almost never the case that local people abruptly switch languages at the border. Instead, they may speak unique dialects that are not easily categorized as one language or another, they may mix elements of more than one language, or they may speak both languages, or even several dialects of each language. Hence the notion of a continuum of dialects, which affects most if not all of the language families in the world, including Romance, Germanic, Slavic, Semitic, Indic, and Sino-Tibetan languages.

DIGLOSSIA

Not all languages, however, exhibit the same degree of variation. Spanish, for example, while geographically dispersed (from Spain, to Latin America, to the Philippines), remains a remarkably unified and consistent language across its regional dialects. As the authoritative survey of the languages of the world[1] puts it, for Spanish "the range of variation is not very great and only rarely disrupts mutual comprehensibility." Arabic, on the other hand, exhibits an extreme case of language variation called diglossia, where a "high" form of the language (literary and prestigious modern standard Arabic) coexists with "low" versions of colloquial Arabic. Modern standard Arabic (MSA), a modernized form of Classical Arabic, is taught in schools as the standard written language but is only spoken by the educated classes in formal situations. What people speak normally are the colloquial versions of the language. These highly varying local dialects, numbering in the hundreds if not thousands, differ in pronunciation, vocabulary, grammar, and style from MSA, to the point of being mutually unintelligible ("one could assemble dozens of Arabs in a room who have never been exposed to the classical language and no one could properly understand each other") and far removed from MSA ("the differences between many colloquial dialects and the classical language are so great that a peasant who had never been to school could hardly understand more than a few scattered words and expressions in it without great difficulty").

PROBLEMS WITH SPEECH-TO-TEXT SYSTEMS

Speech-to-text transcription systems are based on word-based language models, which compute information about which words tend to co-occur. In order to generate such word-based models, large volumes of written text are required, except that, as pointed out in the previous section, textual corpora do not exist for many dialects. Even when a language model can be computed, it is inadequate for dealing with dialect continua, since the way a concept is expressed in one dialect may be quite different in another dialect, even though they are likely to have many words in common. A language model based on a particular dialect is, at best, only useful for that dialect.

Even within the same dialect region, however, the situation is further complicated by the fact that people speak in different registers, which are varieties of language for specific purposes or settings, ranging from very formal to colloquial. In fact, making a finer analysis than simply "high" and "low" versions of Arabic, the eminent Egyptian linguist El-Said Badawi categorizes the registers of Arabic into five different levels[2]: Classical Arabic, modern (MSA), educated colloquial, literate colloquial, and illiterate colloquial. Sometimes people can even mix registers or dialects for stylistic purposes or sociocultural pressures. An example from Cairene Arabic (the dialect of Arabic emanating from the Egyptian capital) is the phrase "he/it is not accepted/admitted" (e.g., to a hospital, university, etc.), which can be expressed in different ways depending on the dialect and register[3]:

la: yuqbalu	(Modern Standard Arabic)
la: yuqbal	(Colloquial form of MSA)
ma: byuqbaʃ	(Mixture of Cairene Arabic and MSA)
ma: byitʔabalʃ	(Colloquial Cairene Arabic)

In addition, when speaking to people from a different dialect, speakers will modify their speech in order to facilitate communication, that is, speakers may switch from one register or dialect to another mid-conversation, a phenomenon that language models are unable to handle. Language models also have difficulty dealing with noise, non-speech sounds, and disfluencies, such as false starts and hesitations. If the speech surrounding a target word or phrase is unclear, there will not be enough context for a language model based system to recognize the target word, even if it is spoken clearly.

PHONETIC-BASED SPEECH ANALYTICS

Because Nexidia's approach to speech analytics is phonetic based, it tends to be more robust against dialectal variation. The phonetic based approach means that Nexidia's speech analyzer uses phonemes, the smallest unit of sound in speech used to make meaningful contrasts between words. Instead of a language model, Nexidia produces an acoustic model to represent the language's phonemes. Thus, when you search for a word or phrase using Nexidia's software, it does not search for the words in a set of transcripts; instead, it searches for the phonemes used to pronounce the words in the acoustic model. Since most languages have only a few dozen phonemes (as opposed to tens of thousands of words), this gives the system more flexibility to locate words or phrases that are not always pronounced the same, even if the acoustic model has not been trained on all the possible variants.

One way in which Nexidia is able to do this is through the use of its patented Interactive Pronunciation Optimization[2], in which the system modifies the pronunciation of the search query based on user feedback. This allows the user to locate words in speech from an unfamiliar dialect by starting with an approximate pronunciation and moving closer to the actual pronunciation with successive iterations. Interactive Pronunciation Optimization is also useful for users who are not fluent in the target language, since it is robust against inaccurate queries.

Suppose, for example, one wished to search for the name of the Palestinian politician *Mahmoud Abbas* in Modern Standard Arabic. A novice user, not fluent in Arabic, might enter *Mahmoud Abbas* as the search query. Initially, Nexidia would represent this query phonetically as *m a **h** m u: d pause **ʔ** i b b a s*, with the ***h*** representing a glottal fricative and the symbol ***ʔ*** representing a glottal stop. In Arabic, however, the 'h' in Mahmoud is actually pronounced as a voiceless pharyngeal fricative (***ħ***) and Abbas begins with a voiced pharyngeal fricative (***ʕ***), two sounds that are difficult for non-native speakers to hear.

Applying Nexidia's Modern Standard Arabic pack to a set of audio recordings, the search term as entered, *Mahmoud Abbas*, results in some true hits but a low overall precision. However, just after two iterations of Interactive Pronunciation Optimization, the system converges to the optimal pronunciation (*m a **ħ** m u: d pause **ʕ** a b b a: s*) and returns hundreds of true hits, achieving a high precision of 92%. The same technique is equally useful for finding words in unfamiliar dialects, such as the Cairene Arabic examples given above, finding foreign words (e.g., the name of a city or person, which may include phonemes not native to the language), or for finding colloquial forms of words using standard vocabulary. Interactive Pronunciation Optimization is also helpful when it is difficult to obtain data from a specific group of people or geographical area.

Phonetic-based speech analytic systems are also more appropriate for languages without a standard writing system. For a speech analytics system based on transcriptions to work, a new orthography must first be created, and all users would have to be trained in this orthography. However, dialectal variation makes this approach unfeasible in many cases, since an orthography based on one dialect may not be adequate for another dialect. Moreover, a fixed set of transcripts would not allow for any deviation from the orthography created for the language. Nexidia's phonetic-based system allows users to enter words as they are pronounced, either by entering the phonemes directly or using a modified version of the orthography of a related language. Interactive Pronunciation Optimization can then be used to refine the search and obtain the most accurate results.

TECHNOLOGY FOR HANDLING LANGUAGE VARIATION

Nexidia recognizes that variation is an integral part of all languages, not just a source of random noise. That is why Nexidia is committed to developing the best technology to help users find what they are looking for, despite the numerous ways that words can be pronounced and ideas can be expressed. Rather than attempting to produce a fixed, word-based transcript that relies on rigid and ultimately inadequate language models, Nexidia's phonetic approach provides the flexibility necessary to accommodate the variation inherent in all human speech.

[1] Comrie, Bernard (ed.). *The World's Major Languages (2nd Edition)*. Routledge, 2009.

[2] Badawi, El-Said. *Mustawayat al-'Arabiyah al-mu'asirah fi Misr* ("Levels of contemporary Arabic in Egypt"). Dar al-Ma'arifah, 1973.

[3] Holes, Clive. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press, 2004.

[4] Morris, Robert *et al*. Wordspotting System. U.S. Patent 7,640,161. Issued December 29, 2009.

Nexidia – Headquarters +1 (866) 355 1241
3565 Piedmont Road NE, Building Two, Suite 400, Atlanta, GA 30305, USA

Nexidia – UK +44 (0)20 8973 2440
Gainsborough House, 2 Sheen Road, Richmond TW9 1AE, United Kingdom