Linguistic Categories for ASR

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- The set of feature vectors extracted from speech signal is a representation of the linguistic categories (or sequence of categories) in the speech.
- These linguistic categories serve as the modeling units for ASR.
- During **training**, these feature vectors are associated with the known category. During **testing**, they are integrated over time to find the best sequence of linguistic categories.

Words

- Words appear to be a natural unit for ASR, both acoustically and linguistically.
- Using words allows the context of phones to be covered in the model.
- In small-vocabulary tasks, this is a good design as there are many instances for each word.
- However, word modeling ignores the commonality of phones in different words. In a large-vocabulary system, this will cause data sparsity.

Phones

- The issue of data sparsity leads to the modeling of sub-word units. The phones are a natural choice.
- Modeling phones is more flexible than modeling words. New word models can be constructed using basic phone models.
- Phone context can be modeled by context-dependent phone models.

Phones vs Phonemes

- The notions of phone and phoneme are often confused.
- Basically, phoneme is an abstraction and phone is an instantiation (of phoneme).
- Allophones: the different phones for a phoneme. For example, aspirated $[p^h]$ and unaspirated $[p]$ correspond to the same phoneme $/p/$.

Phonetic Alphabets

- A phonetic alphabet represents one sound with one symbol.
- A well-known example is the international phonetic alphabet (IPA).
- IPA has a base of about 75 consonants and 25 vowels, covering most languages.
- There is a large inventory (50 or so) of diacritics to modify base phones to achieve finer distinctions.
- For machine readability, an ASCII symbol set (alphabet) is used for sounds, such as the TIMIT alphabet.

Articulatory Features

- We know that some phones are close to one another while others are quite different.
- From a phonetic alphabet, it is not possible to tell which is close to which.
- The articulatory features describe how sounds are pronounced. They can be used to capture the similarity of sounds.

Consonants

- Consonants are made by constricting the tube of vocal tract in various ways, usually with the tongue.
- Two main features for consonants are the place and manner of articulation:
  - The place of articulation is the point of closest constriction in oral cavity.
  - The manner of articulation refers to the amount of constriction in a consonantal gesture.

Place of Articulation

The places of articulation found in English are as follows.

- Bilabial (or labial): [p] [b]
- Labiodental: [f] [v]
- Inter-dental (or dental): [θ]
- Alveolar: [t] [d] [s] [z]
- Palatal/Palatal-alveolar: [$\int$]
- Velar: [k] [g]
- Glottal: lo$\underline{t}$us, ki$\underline{tt}$en

Manner of Articulation

The manners of articulation seen in English are as follows.

- Stop (or plosive): airflow is completely stopped for a short period of time, e.g., [t] [d] [p] [b] [k] [g]
- Nasal: [m] [n]
- Fricative: airflow is constricted but not cut off, for example, [s] [z]
- Affricate: stop + fricative, [t$\int$]
- Liquids and glides: [l], [r], [y]

Vowels

Three parameters tend to be used for vowel articulation.

- Frontness (or backness): refers to the place of largest constriction. For example, [i] is a front vowel, “schwa” is central, and [u] is a back vowel.
- Height: refers to the distance between the jaws. For example, [i] is a high vowel, where the upper and lower jaws are close. [ϵ] is a low vowel.
- Roundness: indicates whether the lips have been rounded. For example, [u] is rounded and [i] is unrounded.

Why Use Features

- Phones of similar articulatory features are similar acoustically.
- We can group phones by the articulatory features. This allows generalization from single phones to phone classes.
- Phones regularly vary in different acoustic contexts. This can be captured by the so-called phonological (or pronunciation) rules.
- Pronunciation rules can be generalized with the articulatory features.

Subword Units

- Subword units refer to the basic units from which the pronunciation of words can be constructed.
- The TIMIT has 61 phones. The CMU dictionary only uses 40 phones. The trade-off is using fewer phones makes the discrimination easier, while using more phones allow finer distinction which may be required for contextual dependency.
- Some systems even use data-driven subword units. The IBM uses clustering to derive “fenones”.

Context-Dependent Phones

- The phonemic categorization leads to units that use less context information.
- The context of a phone can be explicitly modeled.
- Figure 23.2 shows an example of monophone and triphone models.
- Note that the number of models could be an issue for data sparsity. In practice, models can be clustered or model parameters can be tied to alleviate this problem.

Syllables

- In some language, including Mandarin, the syllable is a natural choice.
- Syllable is divided to [onset], nucleus and [coda].
- Simple syllabic types and timing patterns are shown to represent most fluent conversational speech.
- Word boundaries and syllabic boundaries do not have to coincide in fluent speech. Onsets are preferred over codas in English.

Issues in Phonological Modeling

- The pronunciation variability increases dramatically from read-speech (RM, WSJ) to spontaneous speech (switchboard, callhome) recognition tasks.
- The big challenge is to predict such variability.
- Specifically, variability is wildest for
  - speaking rates (correlate to WER)
  - function words (60 per word)
- Pronunciation modeling is important in spontaneous speech. It is a hot and difficult area.