Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 212–232

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Articulatory feature-based pronunciation modeling☆

Karen Livescu [a,*], Preethi Jyothi [b], Eric Fosler-Lussier [c]

[a] *Toyota Technological Institute at Chicago, Chicago, IL, USA*
[b] *Beckman Institute, University of Illinois at Urbana-Champaign, Champaign, IL, USA*
[c] *Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA*

Received 19 July 2014; received in revised form 25 May 2015; accepted 6 July 2015
Available online 16 July 2015

Abstract

Spoken language, especially conversational speech, is characterized by great variability in word pronunciation, including many variants that differ grossly from dictionary prototypes. This is one factor in the poor performance of automatic speech recognizers on conversational speech, and it has been very difficult to mitigate in traditional phone-based approaches to speech recognition. An alternative approach, which has been studied by ourselves and others, is one based on sub-phonetic features rather than phones. In such an approach, a word's pronunciation is represented as multiple streams of phonological features rather than a single stream of phones. Features may correspond to the positions of the speech articulators, such as the lips and tongue, or may be more abstract categories such as manner and place.

This article reviews our work on a particular type of articulatory feature-based pronunciation model. The model allows for asynchrony between features, as well as per-feature substitutions, making it more natural to account for many pronunciation changes that are difficult to handle with phone-based models. Such models can be efficiently represented as dynamic Bayesian networks. The feature-based models improve significantly over phone-based counterparts in terms of frame perplexity and lexical access accuracy. The remainder of the article discusses related work and future directions.

Keywords: Speech recognition; Articulatory features; Pronunciation modeling; Dynamic Bayesian networks

1. Introduction

Human speech is characterized by enormous variability in pronunciation. Two speakers may use different variants of the same word, such as *EE-ther* vs. *EYE-ther*, or they may have different dialectal or non-native accents. There are also speaker-independent causes, such as speaking style – the same words may be pronounced carefully and clearly when reading but more sloppily in conversational or fast speech (Johnson, 2004) (e.g. "probably" may be pronounced "probly" or even "prawly"). In this article we are concerned with building a pronunciation model that is a distribution over the possible sub-word sequences that may be produced in uttering a given word; and we focus on building a model that is as accurate as possible for conversational speech. Here we address speaker-independent pronunciation variability,

☆ This paper has been recommended for acceptance by Katrin Kirchhoff.
* Corresponding author. Tel.: +1 773 834 2549; fax: +1 773 834 9881.
E-mail address: klivescu@ttic.edu (K. Livescu).

http://dx.doi.org/10.1016/j.csl.2015.07.003
0885-2308/

Nomenclature

ASR	automatic speech recognition
IPA	International Phonetic Alphabet
LIP-LOC, LL	lip constriction location
LIP-OPEN, LO	lip opening degree
TT-LOC, TTL	tongue tip constriction location
TT-OPEN, TTO	tongue tip opening degree
TB-LOC, TBL	tongue body constriction location
TB-OPEN, TBO	tongue body opening degree
VELUM, V	velum opening degree
GLOTTIS, G	glottis opening degree
CPT	conditional probability table

i.e. variability due to speaking style or context, although the methods we describe are applicable to studying dialectal or idiolectal variation as well. There are many possible applications for this work, including in automatic speech recognition (ASR), linguistics, and psycholinguistics. In this work, we are mainly motivated by the ASR application, where pronunciation variation in conversational speech is a significant problem.

Pronunciation variation has long been considered a major factor in the poor performance of ASR systems on conversational speech (Ostendorf, 1999; McAllaster et al., 1998; Weintraub et al., 1996a; Fosler-Lussier, 1999; Strik and Cucchiarini, 1999; Livescu et al., 2012). Early work on this topic analyzed this effect in various ways. For example, Weintraub et al. (1996a) compared the error rates of a recognizer on identical word sequences recorded in identical conditions but with different styles of speech, and found the error rate to be almost twice higher for spontaneous conversational sentences than for the same sentences read by the same speakers in a dictation style. Fosler-Lussier (1999) found that words pronounced non-canonically are more likely than canonical productions to be deleted or substituted by an automatic speech recognizer. McAllaster et al. (1998) generated synthetic speech with pronunciations matching canonical dictionary forms, and found that it can be recognized with error rates about eight times lower than for synthetic speech with the pronunciations observed in actual conversational data.

Considering recent advances in speech recognition, one may wonder whether this is still a challenge. Indeed it is: Although error rates in general have gone down dramatically, they are still 50% higher for non-canonically pronounced words in a recent discriminatively trained recognizer (Livescu et al., 2012). Most speech recognition systems use context-dependent (such as triphone) acoustic models to implicitly capture some of the pronunciation variations. However, this approach may be insufficient for modeling pronunciation effects that involve more than a single phone and its immediate neighbors, such as the rounding of [s] in *strawberry*. The work of Jurafsky et al. (2001) suggests that triphones are in general mostly adequate for modeling phone substitutions, but inadequate for handling insertions and deletions.

There have been a number of approaches proposed for handling this variability in the context of phone-based speech recognition. One approach, which was studied heavily especially in the 1990s but also more recently, is to start with a dictionary containing canonical pronunciations and add to it those alternative pronunciations that occur often in some database, or that are generated by deterministic or probabilistic phonetic substitution, insertion, and deletion rules (e.g., Sloboda and Waibel, 1996; Riley et al., 1999; Weintraub et al., 1996b; Strik and Cucchiarini, 1999; Fosler-Lussier, 1999; Saraçlar and Khudanpur, 2004; Hazen et al., 2005). Other approaches are based on alternative models of transformations between the canonical and observed pronunciations, such as phonetic edit distance models (Hutchinson and Droppo, 2011) and log-linear models with features based on canonical-observed phone string combinations (Zweig and Nguyen, 2009). Efforts to use such ideas in ASR systems have produced performance gains, but not of sufficient magnitude to solve the pronunciation variation problem.

One often-cited problem is that with the introduction of additional pronunciations, confusability between words is also introduced (Finke and Waibel, 1997; Riley et al., 1999). This may be due to the large granularity of phone-level descriptions: An actual pronunciation may contain a sound that is neither a dictionary phone nor an entirely different phone, but rather something intermediate (Saraçlar and Khudanpur, 2004), suggesting that a finer-grained level of