Using Articulatory Representations to Detect Segmental Errors in Nonnative Pronunciation

Joseph Tepperman, *Student Member, IEEE*, and Shrikanth Narayanan, *Senior Member, IEEE*

***Abstract*—Motivated by potential applications in second-language pedagogy, we present a novel approach to using articulatory information to improve automatic detection of typical phone-level errors made by nonnative speakers of English—a difficult task that involves discrimination between close pronunciations. We describe a reformulation of the hidden-articulator Markov model (HAMM) framework that is appropriate for the pronunciation evaluation domain. Model training requires no direct articulatory measurement, but rather involves a constrained and interpolated mapping from phone-level transcriptions to a set of physically and numerically meaningful articulatory representations. Here, we define two new methods of deriving articulatory-based features for classification: one, by concatenating articulatory recognition results over eight streams representative of the vocal tract's constituents; the other, by calculating multidimensional articulatory confidence scores within these representations based on general linguistic knowledge of articulatory variants. After adding these articulatory features to traditional phone-level confidence scores, our results demonstrate absolute reductions in combined error rates for verification of segment-level pronunciations produced by nonnative speakers in the ISLE corpus by as much as 16%–17% for some target segments, and a 3%–4% absolute improvement overall.**

***Index Terms*—Articulatory features, hidden-articulator Markov model (HAMM), language learning, nonnative speech, pronunciation evaluation, reading assessment.**

I. Introduction

PRONUNCIATION evaluation is the discrimination of categories of pronunciation based on the variability expected in speech. Whereas traditional speech recognition is concerned with identifying the features common to all spoken realizations of a phoneme, word, or phrase, the automatic pronunciation evaluation task intends to detect the often subtle contrast between one realization and another, and to quantify this difference. This calls for the establishment of a reference abstract pronunciation model—the canonical form—against which all realizations can be compared, and naturally lends itself to applications in the area of second-language acquisition, in which a student's pronunciation will be assessed relative to a "gold standard." With these pedagogical applications in mind, for nonnative speakers we define our canonical reference to be what linguists often call the single "citation form" [1]—the formalized lexical pronunciation of words when spoken in isolation, which students of a foreign language are expected to produce when practicing their new tongue.

The expected deviants from the canonical form can be modeled in a number of ways and on various time-scales, traditionally referred to as the segmental, suprasegmental, and articulatory feature levels. A segment-level pronunciation error we define here as the substitution of a phoneme or sequence of phonemes, with respect to the canonical form. A phoneme is an abstract but unambiguous unit (or "segment") of speech, and the term phone is given to a specific realization of that unit; by definition, a sound is a phoneme if simply changing that sound can change the meaning of a word [1]. In English, certain events within a phone will not change the meaning of the word they compose nor the phonemes that make up that word, so these variants—called "allophonic"—necessarily occur below the segmental level. Pronunciation mistakes in the subphonemic or articulatory feature scale are those which result in a segment- or word-level substitution; for example, a lengthy voice onset time (VOT) in the consonant /d/ can lead to the perceived substitution of /t/ [1], [2]. Suprasegmental pronunciation errors may encompass several phonemes and involve longer-scale variations in prosody, including those of syllabic or word-level stress [3]. The set of expected errors may be arbitrarily large, of course. To constrain the search space, past work in modeling pronunciation variants or errors has derived these expected forms with rule-based or statistical methods of transforming the canonical representation. The rule-based methods [4], [5] are grounded on firm linguistic theory and expected usage, and allow for a generalized and unsupervised approach to future, unseen datasets, but statistical analysis [6]–[8] can estimate probabilistic models for each rule's application, and will sometimes show a certain rule to be statistically insignificant for specific cases.

In this paper, we choose to focus on segmental pronunciation errors and, more specifically, those which can be modeled as systematic—i.e., phonologically or statistically predictable—based on prior knowledge of the speaker's native language and the data set. Our goal is to discriminate between a canonical segment and any of its substitutions in nonnative speech, to disambiguate between a close approximation of the canonical form and a true pronunciation error. It is the systematic substitutions—the ones a speaker produces unconsciously by the "phonological transfer" from his native tongue—which

Manuscript received December 22, 2006; revised July 25, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Simon King.

The authors are with the Integrated Media Systems Center, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: tepperma@usc.edu; shri@sipi.usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.909330

are most difficult to correct and therefore most important to practice [9].

This evaluation of pronunciation on the phone level is essentially one formulation of the traditional hypothesis verification task. In verifying that an observation utterance O fits a target model M_t, we calculate the posterior probability

$$P(M_t|O) = \frac{P(O|M_t)P(M_t)}{P(O)} \tag{1}$$

by approximating $P(O) \approx P(O|M_f)P(M_f)$ where M_f is a general or specific substitution (or "filler") model. If we assume equal priors, this becomes the likelihood ratio (or "confidence score")

$$\tau = \frac{P(O|M_t)}{P(O|M_f)} \tag{2}$$

and we decide in favor of verification if $\tau \geq T$ for some threshold T. Many different types of filler models have been proposed, depending on the application. A typical baseline approach is to make the filler be a generalized "garbage" model for all speech, either on the phone or word level, though a specific set of "cohort" models can improve these scores' reliability [10]–[13]. In the domain of speaker identification and speaker verification, a complex set of "impostor" models is used as the discriminative filler [14].

In the past, the detection and correction of pronunciation errors on the phone level has relied on this general approach to verification and scoring, given a set of trained phoneme models [7], [8]. However, even when performing verification on the segment level, a strictly phonemic representation might not tell the "whole story" about the nature of a systematic mistake. Consider the common substitution of /s/ for /z/ made by native German speakers in such English words as "dessert" and "warnings," as predicted by German orthography and phonology [9]. Though distinct phonetic units, both phonemes in question share a common Place and Manner assignment in terms of articulation—alveolar fricative—and differ only in that /z/ is voiced and /s/ is not. This type of "close" substitution, this overlap between the canonical form and its common substitutions once factored into the articulatory domain, is the rule rather than the exception when it comes to nonnative speech. To model such an error in terms of parameterized articulatory motion has more explanatory power than to treat it as a simple substitution. We begin this study with the hypothesis that the sorts of insights offered by articulatory information will allow us to discriminate between the canonical form and its close segmental substitutions in nonnative speech more astutely than with the more ambiguous phonemic models alone.

Articulatory representations of speech have been used to improve accuracy in such tasks as speaker verification [15], general speech recognition [16], [17], pronunciation modeling [5], and spectrally impoverished or whispery speech recognition [18]. However, direct articulatory measurement and transcription is not always available (especially for nonnative speech), and not all articulatory representations are suitable for the pronunciation evaluation task. In this paper, we present a rule-based method for deriving useful articulatory representations from phoneme-level transcriptions. With these representations, we train models for our articulatory derivations as well as standard phoneme models using the same traditional spectral features. We then show how these models can be used to generate novel features and confidence scores for segment-level pronunciation verification. We intend to show that features derived from an articulatory representation, when combined with traditional phonemic features, will improve verification accuracy in phone-level evaluation. Moreover, we approach it with the dearth of human input demanded by automated language-learning applications.

Section II presents the corpora used in this study, and in Section III, we give some background on work in articulatory representations, and explain our own chosen representation. Section IV describes the model training procedure for these representations, while Section V outlines several new methods for deriving features from these models. Section VI explains the pronunciation evaluation experiments performed, based on these features, and Section VII discusses the results. In Sections VIII and IX, we conclude and present some ideas for future work in this area.

This is an expansion and reformulation of similar work reported previously in [19].

II. Corpora

The data used in our experiments were compiled by the University of Leeds in their ISLE corpus [9]. These recordings consist of 46 adult Intermediate British English learners who are native speakers of either Italian or German—23 of each. Utterance prompts were complete sentences designed to highlight specific difficulties English learners typically encounter in pronouncing single phone minimal pairs, phone clusters, and primary stress minimal pairs. The recordings were automatically segmented by a forced-aligner. These transcriptions were then augmented on the phone level by a team of five linguists to reflect each speaker's pronunciation. However, no effort was made to correct discrepancies in the automatic segmentation times. Consequently, this means that the derivation of our features is fully automated, though the test set annotations provide a reference for assessing our method's performance.

For comparison with the nonnative results, additional experiments on native speakers were done using the MOCHA-TIMIT corpus [20] and the IViE corpus [21], both of which are composed of read British English of the sort the ISLE corpus students were learning to speak—we used 43 Southern British speakers in all. Though the pronunciation of native speakers does not diverge from the canonical in the same way as that of nonnative speakers, experiments on these corpora can legitimize our proposed features for assessment of native speech, in which the usefulness of articulatory information has already been well documented elsewhere.

III. Articulatory Representations

A. Previous Work

Let us begin by defining some terms. By an *articulatory representation* we mean any convention of spoken language transcription that uses symbols denoting an abstraction of an underlying speech production mechanism or position (as opposed to symbols that represent abstractions of perceptual or semantic phenomena). Often, an articulatory representation will span multiple *streams*, which is to say that the symbols used in the representation can be grouped based on their relevant components. We refer to the symbols that compose these streams as articulatory *classes*. One example stream could be the Manner of articulation, and its classes may include Fricative, Stop, and Vowel. Another stream could be the degree of Lip Rounding, with classes of perhaps Rounded, Neutral, and Spread.

Many variations on the idea of representing speech through an articulatory framework already exist, though the fundamental methodology has relied either on articulatory measurement, or a mapping to the articulatory feature domain from a higher-level representation (ordinarily that of phonemes). Electromagnetic or electropalatal measurement as used in [22] is sometimes costly, though still appealing for the pronunciation evaluation task because of the concrete physical referents for any models derived therefrom—it allows us to "point to" the observable differences in production among realizations, and we may make our class assignments as specific as the resolution of our vocal tract imaging permits. As yet, no known corpus of direct articulatory measurement possesses the same scope of nonnative variability as the ISLE recordings.

As for the rule-based approaches to generating an expected articulatory representation from phonemic transcriptions, several class and feature configurations have been proposed, all grounded in articulatory feature theory [1]. The feature assignments defined by Kirchoff in [16]—Voicing, Manner, Place, Front-Back, and Rounding—have served as an appropriate baseline for the experiments reported in [15], [18], and [23], but are perhaps too abstract and coarse to be of much use in pronunciation evaluation or language learning applications. An ideal articulatory representation for this task should allow for differences between vowels subtler than simply Low or High, Rounded or Unrounded, and so on. In this paradigm, the model configurations for certain pairs of often substituted vowels are not always distinct, making them theoretically impossible to classify automatically. A good example is the substitution of /i:/ for /ɪ/ frequently made by Italian learners of English—[16]'s feature space renders them both as high, front vowels, without distinction. In addition, all classes within a given articulatory stream should, when taken as a group, delineate a graduated physical progression within that stream. This is necessary to ensure that the models derived from each stream will represent pronunciation-dependent variations in articulatory motion over time; we should be able to visualize the overlapping motion of each vocal tract component, so as to parameterize the unique qualities of each pronunciation realization effectively. This is not possible, though, if we treat all abstract Place models as members of the same stream, as in [16]; a Labial articulation concerns the lips, whereas a Velar articulation concerns the tongue and soft palate—they should be evaluated and tracked separately. As another example, if an abstract Manner classifier requires a mutual exclusion between Nasal and Vowel classes (as [16] has done), this disallows the possibility of classifying a nasalized vowel into both categories.

Acoustic modeling of an articulatory representation can be done using traditional spectral features—cepstral coefficients, for example—and any statistical classifier. Previous studies have shown success in articulatory modeling using neural networks [16]. Hidden-articulator Markov models (HAMMs) were first proposed by Richardson *et al.* in [17] as a method of incorporating articulatory information into the existing hidden Markov model framework that dominates traditional speech recognition. HAMMs have been used to improve speech recognition performance when used in combination with traditional HMMs, and also show a robustness to noise previously unseen in phoneme-based acoustic modeling [24]. This is remarkable in light of the fact that both HAMMs and phonemic HMMs are trained on the same spectral features, so it is really just the augmented representation which is responsible for this improvement in model performance.

A hidden Markov model consists of j states, each with an output density $b_j(o_t)$ modeling the probability that state j "generates" the speech observation o_t, and a set of transition probabilities such that a_{ij} models the probability of a transition from state i to state j [25]. We assume that, given a model M, the joint probability of a sequence of observations $O = o_1, o_2, \ldots, o_T$ and their underlying "hidden" sequence of states $X = x(1), x(2), \ldots, x(T)$ can be calculated as the product of all output and transition probabilities over that sequence, as follows:

$$P(O, X|M) = \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)}. \tag{3}$$

The state sequence is not in general known, so we must compute an observation's likelihood given a model by summing over all possible state sequences, as follows:

$$P(O|M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \tag{4}$$

where $x(0)$ and $x(T+1)$ are model entry and exit states, respectively. In the case of an HAMM, the hidden state sequence represents changes in the speech spectrum over time within a particular articulatory class or vector of such classes, rather than over the more abstract phoneme units ordinarily represented by HMMs.

Each HAMM state proposed in [17] represents what is called an *articulatory configuration*—a vector of integers $C = \{c_1, c_2, \ldots, c_N\}$ over N streams where $0 \leq c_a < M_a$ and M_a is the cardinality of stream a. A similar "feature-bundle" state was proposed by Sun and Deng in [4] but without the numerical mapping. This integer representation is advantageous for several reasons. It treats each stream as a set of discrete

TABLE I
OUR ARTICULATORY FEATURE SPACE. NOTE THE NUMERICAL MAPPING AND GRADUAL PHYSICAL PROGRESSION AMONG CLASSES WITHIN A GIVEN STREAM, PROPORTIONAL TO THE INTEGERS CHOSEN TO REPRESENT THOSE CLASSES

stream	*classes*	*cardinality*
jaw	0: Nearly Closed, 1: Neutral, 2: Slightly Lowered, 3: Lowered	4
lip separation	0: Closed, 1: Slightly Apart, 2: Apart, 3: Wide Apart	4
lip rounding	0: Rounded, 1: Slightly Rounded, 2: Neutral, 3: Spread	4
tongue frontness	0: Back, 1: Slightly Back, 2: Neutral, 3: Slightly Front, 4: Front	5
tongue height	0: Low, 1: Mid, 2: Mid-High, 3: High	4
tongue tip	0: Low, 1: Neutral, 2: Dental, 3: Nearly Alveolar, 4: Alveolar	5
velum	0: Closed, 1: Open	2
voicing	0: Unvoiced, 1: Voiced	2

articulatory classes, isomorphically mapping numerical values to physical positions for factored tracking of movement over time. This permitted [17] to mathematically impose a set of static and dynamic constraints on state transitions, so that the streams could move asynchronously but still adhere to the physical properties of the vocal tract.

In [17], the articulatory representation was derived from phone-level transcripts such that some configurations necessary for constrained transition would not be seen in the training data—i.e., no phoneme mapped to many configurations which were still physically possible and were needed for modeling the transitions between phones. The integer representation allowed [17] to interpolate the models for these meta-phonemic states as the Cartesian product of the phoneme-trained states. This required the estimation of several hundred thousand HAMM parameters—one disadvantage of modeling a complete configuration vector rather than the individual streams.

One related approach, implemented by Livescu and Glass in [5] and Wester *et al.* in [26], is to model the articulatory representation as a dynamic Bayesian network in which each hidden state is factored into its respective streams, allowing for trainable probabilistic dependencies among them. As a modification to the original HAMM definition stated above, this means that $P(O, X|M)$ can be calculated as the product of output and transition probabilities as well as the probability of asynchronous movement among the parameterized states in X. For speech recognition purposes, this is an improvement over the approach in [4], [17] because it not only allows for fixed constraints in allowable articulatory configurations, but succeeds in modeling the relative probability of one configuration over another.

B. Choice of Representation and Modeling

Chosen because of their fine-grained taxonomy in concrete physical terms, the articulatory representation used in this study (enumerated in Table I) is based primarily on the

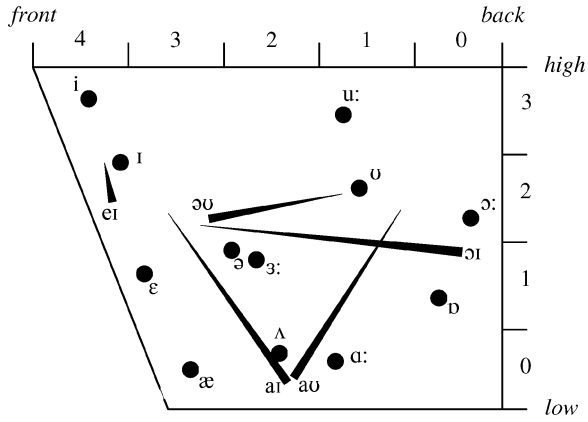

Fig. 1. British English (BBC newscaster) vowel chart with numerical mappings, after [1]. This illustrates the relationship between the tongue frontness and tongue height streams.

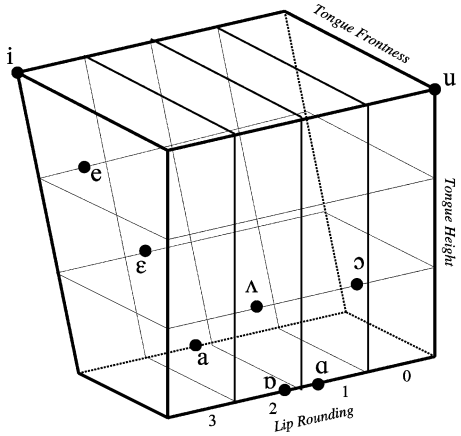

Fig. 2. A 3-D representation of the relation between the lip rounding, tongue frontness, and tongue height streams in cardinal and secondary English vowels, after [1]. Think of this as a rotation of Fig. 1, to reveal the lip rounding dimension. This also explains the numerical mapping assignments for the lip rounding stream.

mapping and modeling proposed in [17], but with some important differences. First, we make many modifications to [17]'s phone-to-articulator mapping, based on our interest in modeling speech that is specifically nonnative. The students in the ISLE corpus were learning British English, so most of the mappings in [17] for American English do not serve as appropriate reference-points—particularly those for the relative tongue positions of the vowels. These we determine anew by drawing from Ladefoged's charts of BBC newscaster English and relative lip rounding of the cardinal and secondary vowels [1], here reflected in Figs. 1 and 2. We also incorporate Ladefoged's rules for allophonic variability in English, but omit many of them because they do not apply to our chosen representation, or are not generalizable to nonnative speakers, or both. For example, the rule that a glottal stop /ʔ/ substitutes for /t/ when preceding an alveolar nasal is omitted because neither our articulatory representation nor the ISLE corpus' phoneme set accounts for the glottal stop; furthermore, Ladefoged suggests this rule applies to "many"—but not all—accents of English. The final set of four contextual rules we used are as follows.

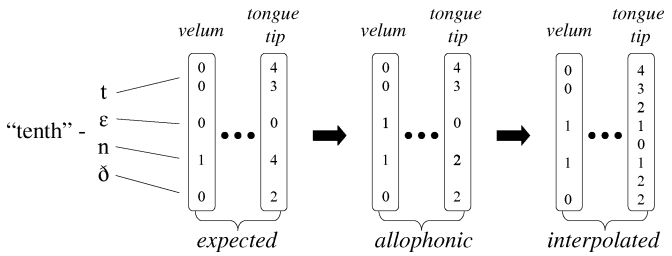

Fig. 3. Illustration of the three steps involved in automatically deriving an articulatory-level representation from phone-level transcriptions, as explained in Section III-B.

- A vowel preceding a nasal consonant will also be nasalized.
- Voiced stops and affricates become unvoiced when syllable initial.
- Stops are unexploded when occurring immediately before another stop.
- Alveolar consonants become dental in anticipation of a subsequent dental consonant.

The algorithm for expanding phone-level transcriptions to a sequence of expected articulatory representations involves three steps. First, a phone is mapped directly to its expected articulatory configuration, in eight dimensions; charts of this mapping, derived from [1] and [17] are given in the Appendix. Then, based on its context, the expected configuration is perhaps changed to one of its allophonic variants, based on the above rules. Finally, the overall transcript for the entire utterance is interpolated so as to adhere to the physical constraints of the vocal tract. No model states may be skipped in the transcripts' numerical transitions (unless transcribed Silence intervenes) because they represent a discrete sequence of positions. For example, a transition of the Tongue Frontness stream from Back (class 0) to Front (class 4) must first pass through all intermediate classes (1 through 3), because that is what a real tongue would do. In this way, without the need for vocal tract imaging, we can generate a series of articulations for a phone sequence, for use as an abstract standard of expected behavior from which all model parameters for this articulatory representation may be derived. This representation is ultimately synchronous with the original phoneme labels, though the models derived from them need not perform synchronously. In a sense it is also noncausal because the present articulatory class is always dependent on future contextual information. See Fig. 3 for a graphical depiction of our transcription expansion technique.

Our modeling method also owes a debt to the HAMMs of [17] but, like the method of representation, the models have been adapted to suit the domain of nonnative speech evaluation. Rather than treating each possible configuration of articulatory classes as a unique hidden state—something of a "metaphoneme"—we designed a separate set of hidden-articulator Markov Models for each of the eight streams (Jaw, Lip Separation, etc.), and trained each set independently. This allowed for free asynchrony among the articulators, so that the results might mimic the overlapping behavior of a true vocal tract's constituent parts. It also permitted the generation of independent bigram models specific to each stream, and simplified the training and testing process, since in this quantization scheme no feature has more than five classes, compared to the several thousand states trained in the previous work of [17].

These simplifications rest on the assumption of independent motion among these eight articulator streams, which in a different study might not be valid—it allows for results that could potentially violate the fundamental physical constraints of the human vocal tract (e.g., dependencies between the jaw position and lip separation, tongue tip and tongue body, etc.). However, the point of this project is not to build an articulator-based speech recognizer or even a general phoneme recognizer, in both of which such constraints would be more important. Rather, we intend to demonstrate improved discrimination between canonical and noncanonical pronunciations in articulatory feature space, regardless of the accuracy in recognizing any of the individual articulatory streams (the true transcripts of which we do not know). Verification in this domain may in fact perform better under an assumption of independence. The true articulation may diverge from its expected mapping, especially in the case of a segmental substitution. Disallowing physically impossible articulatory configurations from the recognition results may limit the representation of fine pronunciation distinctions within articulatory feature space. With well-trained models, if the results point toward a physically unlikely articulation, it could signify the presence of a pronunciation mistake, and that is exactly what we intend to detect.

IV. Model Training

The goal of the experiments in this study is to show that features derived from models trained on an articulatory representation, when combined with features derived from traditional phoneme models, can be used to evaluate a nonnative speaker's segment-level pronunciation more accurately than with segmental features alone. This required us to train traditional phoneme models as well as articulatory-based models, both based on the Block D recordings of the ISLE corpus of nonnative speech, described in Section II. Block D consists of read sentences in both question-and-answer and "I said X, not Y" forms, designed to contrast spellings ordinarily difficult for nonnative speakers. For the Italian and German speakers combined, this Block of recordings totals about 5 h of training speech.

We trained separate phone-level hidden Markov models for the native German and Italian speakers in the ISLE corpus using standard techniques. The first 12 Mel cepstral frequency coefficients (MFCCs) (plus delta, acceleration, and normalized energy coefficients) were extracted every 10 ms using a window size of 16 ms. These features were then used to initialize three-state context-independent HMMs over the ISLE phone set (plus Silence and generalized phone-level filler models), using segmentation times from the automatic alignment provided by the corpus' transcriptions. The models were then refined with embedded reestimation and by updating the number of mixtures per state to 16. Context-dependent models were not trained because

all of our HMM alignment-based features were to be decoded over segments taken in isolation, and not over an entire word or phrase.

Eight different sets of HAMMs—one set for each of the articulatory streams enumerated in Table I—were also trained, again keeping the native German and Italian speakers separate. All model parameters and topology were the same as for the phone models except the training procedure was slightly different. The ISLE corpus does not provide transcripts below the segment level, so we mapped the phoneme transcriptions to a sequence of articulatory classes—paramaterized in eight streams—according to the convention outlined in Section III-B. For this new articulatory expansion, we did not have a reliable segmentation (especially for articulatory motion within a phone) so the HAMMs were initialized using a flat-start procedure, then updated with embedded reestimation.

For purposes of articulatory recognition over an entire utterance, we also trained a bigram articulation model for transitions among articulatory classes in each of the eight streams, based on our interpolated transcripts. A bigram model assumes that the current articulatory position for a given stream depends only on the position immediately preceding it. The purpose of this was to steer the recognizer away from physically unlikely transitions within each of the eight streams, e.g., if the current Tongue Tip recognition result is Low (class 0) then the next recognition result can either stay in that position or move to Neutral (class 1). The interpolated transcripts were also constrained as such, therefore the bigram model derived from them would be as well. And a simple bigram model is sufficient to capture these transition rules.

All model training was done using HTK [25]. Comparable models for native speakers of British English were trained using the Southern speakers of the IViE corpus and half of the MOCHA-TIMIT database, which both consist of phonetically balanced read speech, totaling about 1 h of training data.

V. Derivation of Features

We propose three types of features for this verification problem. The first type are confidence scores derived from likelihoods based on alignment of traditional phone-level HMMs. Though they have been shown to provide complementary information to phonemic models, HAMMs alone do not perform as well as phone models in basic recognition tasks [17]. These HMM scores serve as a tool for both baseline experiments and ones in conjunction with features derived from articulatory models, to be combined in a decision tree framework in Section VI. We divide articulatory-based features into two categories: those based on articulatory recognition results (represented by integers as explained in Section III); and those which are confidence scores for the expected articulation in each of the eight streams, directly analogous to the alignment-based HMM scores. We explain details about these feature sets below.

TABLE II
Relative Substitution Probabilities for Commonly Mispronounced Phonemes, for the ISLE Corpus' Native German Speakers. The Probabilities for a Given Target May Not Add Up to 1 Because All Substitutions of Less Than 0.01 Probability Have Been Disregarded

target	*substitution*	*probability*	*target*	*substitution*	*probability*
v	f	0.76	ə	ʊ	0.21
	w	0.21		ɒ	0.18
t	SIL	0.73		uː	0.12
	d	0.21		æ	0.11
z	s	0.91		ɛ	0.07
	ɪz	0.01		əʊ	0.07
	ɪs	0.01		ɪ	0.06
	SIL	0.01		eɪ	0.04
ʌ	ə	0.9		SIL	0.02
	ɒ	0.02		ʌ	0.02
	ɪ	0.01		ɜː	0.01
	ɔː	0.01			

A. Traditional HMM Confidence Scores

In Section I, we gave some background on pronunciation evaluation using confidence scores estimated from appropriate filler models, for verification of the canonical form. Here we investigate three different HMM-based fillers, based partly on those used in [27], so as to select the best of these derived confidence scores as our baseline feature.

The best use of our prior knowledge of expected substitutions is to let the filler be a recognition network over all likely substitutions in the target domain for the canonical phone in question, with relative arc transitions derived from corpus statistics. Formally, we define the confidence measure's denominator in (2) to be

$$\max_i \{P(O|p_i)P(p_i)\} \tag{5}$$

where O is the segmental speech observation, p_i is one phoneme model (or sequence of phoneme models) in the set of expected substitutions (i takes all values in this set), and $P(p_i)$ is the prior probability of the substitution p_i. The resulting ratio used as our confidence score is

$$\tau = P(O|p_t)P(p_t)/\max_i \{P(O|p_i)P(p_i)\} \tag{6}$$

where p_t is the target phoneme's model, and we set $P(p_t) = 1$ on the assumption that there is only one target. This particular confidence score, based on expectations of common phone-level substitutions, we will refer to as P_{subs} in the remainder of the paper. For a concise list of substitution statistics used in constructing these filler networks, please see Tables II and III.

The second proposed HMM-derived confidence measure is based on what [27] calls a "phoneme loop" filler, which is identical to that of the denominator in (6), except i can take values over all phone-level HMMs, and $P(p_i) = P(p_j)$ for every i and j in the complete HMM set, i.e., all substitutions are weighted equally. The confidence score derived from this phoneme loop filler we abbreviate as PL from now on.

TABLE III
RELATIVE SUBSTITUTION PROBABILITIES FOR COMMONLY MISPRONOUNCED PHONEMES, FOR THE ISLE CORPUS' NATIVE ITALIAN SPEAKERS. THE PROBABILITIES FOR A GIVEN TARGET MAY NOT ADD UP TO 1 BECAUSE ALL SUBSTITUTIONS OF LESS THAN 0.01 PROBABILITY HAVE BEEN DISREGARDED

target	*substitution*	*probability*	*target*	*substitution*	*probability*	*target*	*substitution*	*probability*
t	tə	0.47	ʌ	ə	0.59	ə	ɒ	0.2
	SIL	0.36		ɑː	0.11		uː	0.14
	d	0.06		ɒ	0.08		ɛ	0.14
	θ	0.01		ɛ	0.03		ʊ	0.1
	də	0.01		ʊ	0.03		æ	0.1
	ə	0.01		uː	0.03		ʌ	0.07
ŋ	ŋg	0.56		ʌp	0.01		əʊ	0.06
	ŋgə	0.26		æ	0.01		eɪ	0.02
	n	0.06		əʊ	0.01		ɪ	0.02
	ŋə	0.05	ɜː	ɜːɪ	0.44		i	0.01
	ŋk	0.02		ɛɪ	0.14		ɑː	0.01
ɪ	i	0.81		ɔːɪ	0.1		SIL	0.01
	ə	0.06		uːɪ	0.07	ʊ	uː	0.68
	ɛ	0.04		ɒɪ	0.03		ʌ	0.14
	aɪ	0.01		ɹ	0.02		uːl	0.05
	æ	0.01		ɔː	0.01		ʊl	0.05
	SIL	0.01		ʊɪ	0.01		əʊ	0.01
	ʌ	0.01		əɪ	0.01		ɑː	0.01
							ɒ	0.01

A third filler we define as a generalized segment-level HMM trained on all phonemes. Its confidence measure is calculated as

$$\tau = P(O|p_t)/P(O|f_g) \tag{7}$$

where p_t is the target phoneme and f_g is the generalized filler.

B. Articulatory Confidence Scores

In a manner directly analogous to the derivation of the HMM alignment scores, we generated articulatory alignment confidence scores over all eight articulation streams by constructing ratios of target to filler likelihoods. However, unlike the P_{subs} HMM-based confidence score, these target and filler models required only a general knowledge of articulatory phonology and no special prior knowledge of the corpus statistics for likely substitutions. We define the target articulatory model for each phone of interest as a recognition network of allowable articulatory classes for that particular phone and stream. For example, all vowels are allowed to decode the Velum stream as either Open or Closed (1 or 0 by the numerical mapping), since in English the difference is not contrastive—a nasalized vowel is only an allophonic variant of that vowel, not an entirely new phonetic unit. We define the filler model against which the target is compared as the recognition network of all unallowable articulation classes within that stream, for the phone in question—the complement of the target network. This includes decoding Silence as one unallowable articulation. No weights are assigned to the arcs of these recognition networks because, unlike the phonemic substitutions, the articulation statistics for the ISLE corpus are not available.

Formally, for a segment-level speech observation O, we define a vector of articulatory confidence scores in N streams, $A = \{\tau_1, \tau_2, \ldots, \tau_N\}$, where

$$\tau_a = \max_i \{P(O|t_i)\} \,/\, \max_j \{P(O|f_j)\} \tag{8}$$

t_i is one of I allowable (target) articulatory classes for this segment, and f_j is one of J unallowable (filler) articulatory classes for this segment. In our case, $N = 8$, and I and J vary depending on the stream and the segment. We include Silence in the filler network, so $I+J = M_a+1$ where M_a is the cardinality of stream a. All scores in vector A can be used as classification features, and together the whole vector of scores will be denoted as A_{conf} in the rest of the paper.

C. Articulatory Recognition

A second novel method of deriving useful features from articulatory models for segmental pronunciation discrimination is as follows. We represent all articulatory classes within each of the eight streams as integers, consistent with the linguistically based method outlined in Section III. The magnitude of these integers is proportional to the discrete positions they represent, so the results of articulatory recognition over these eight streams within a segment of interest can in themselves serve as an eight-dimensional feature vector in articulatory space—an articulatory factorization of that segment.

We propose two methods of deriving these recognition-based feature vectors. First, we argue that articulatory recognition over an utterance (rather than just the isolated segment of interest) using a bigram articulation model is necessary to ensure that we capture the appropriate degree of articulatory motion within a phoneme. The recognition results need not be synchronous

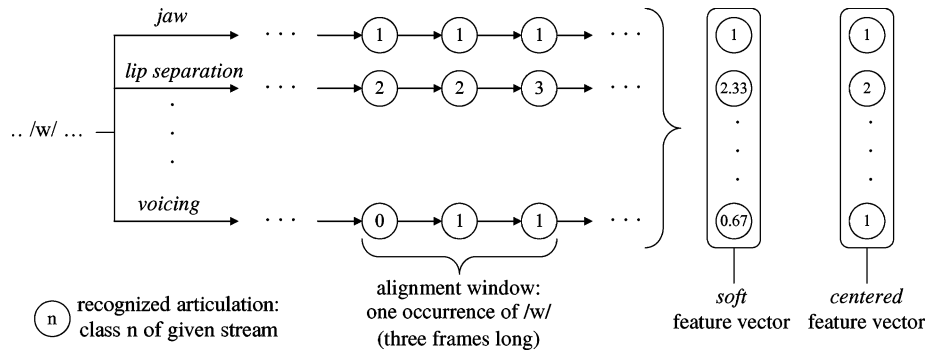

Fig. 4. Example of recognizing eight HAMM streams and then extracting the eight-dimensional *soft* and *centered* articulatory feature vectors from them. This correct realization of the target /w/ will be grouped with others so as to distinguish them from the class of substitution pronunciations.

with the ISLE corpus' phone-level segmentation, which is based on automatic alignment with the canonical pronunciation and should therefore not be blindly trusted on the level of articulation, especially if the pronounced phone differs from the canonical form. If we were to perform recognition only over the interval of each phone of interest, the classifier might miss crucial articulatory transitions immediately before or after the phone's transcribed start and end points. A bigram model captures local context important for classification, and, though the streams are modeled as independent, it ensures that a physically impossible transition is not likely to be made within any of the streams, as explained in Sections III and IV. The articulatory recognition segmentation may not line up perfectly with the phone-level alignment, and may include a sequence of articulatory classes within a given phone's interval. This allows for multiple interpretations of what should be considered the ultimate vector of recognition results. We investigate two such methods, referred to as *centered* and *soft* and explained in detail in the remainder of this section.

For a speech sequence $O = o_1, o_2, \ldots, o_n$ in n frames, the utterance-level articulatory recognition result for stream a with cardinality M_a is given by

$$\begin{aligned}\widehat{C}_a &= \arg\max_{C_a} \{P(C_a|O)\} \\ &= \arg\max_{C_a} \left\{ \frac{P(O|C_a)P(C_a)}{P(O)} \right\} \\ &= \arg\max_{C_a} \{P(O|C_a)P(C_a)\} \end{aligned} \tag{9}$$

where $C_a = c_a^1, c_a^2, \ldots, c_a^m$ is a sequence of m integer articulatory class labels such that $0 \leq c_a^i < M_a$ and $m \leq n$, i.e., these class labels can span more than one frame. Because we are using a bigram articulations model, we compute the approximation

$$P(C_a) = P\left(c_a^1, c_a^2, \ldots, c_a^m\right) \approx P\left(c_a^1\right) \prod_{i=2}^{m} P\left(c_a^i | c_a^{i-1}\right) \tag{10}$$

where each class depends only on the preceding class in its stream. Once recognized, we choose to express $\widehat{C}_a$ as the sequence of integer class labels by frame—$s_a^1, s_a^2, \ldots s_a^n$—with the same frame-level indices as in O. Now, for a segment within this utterance beginning at frame b and ending at frame e, we define the segment-level articulatory recognition result in stream a two ways—

- *centered*:

$$\widehat{S}_a = s_a^{\left\lfloor \frac{(b+e)}{2} \right\rfloor} \tag{11}$$

- *soft*:

$$\widehat{S}_a = \frac{1}{e-b} \sum_{i=b}^{e} s_a^i. \tag{12}$$

Combining the eight streams, one segment's vector of recognition results is $R = \{\widehat{S}_1, \widehat{S}_2, \ldots, \widehat{S}_8\}$. To make this vector target-independent as a set of features, we subtract each stream's recognition result from the expected articulation: $\widehat{R} = \{L_1^t - \widehat{S}_1, L_2^t - \widehat{S}_2, \ldots, L_8^t - \widehat{S}_8\}$ where L_a^t is the expected integer class label for segment t in stream a; if t is expected to have articulatory motion below the phone level, then L_a^t is the mean of all expected integer class labels within t for stream a. The final feature vector $\widehat{R}$ reflects the distance of the recognition result from the expected mapping of the canonical, regardless of the target.

The *centered* method (its features denoted by A_{cent}) is an accepted one for evaluating articulatory recognition results, though it ignores any articulatory motion within a phone. For this reason, other studies that have followed [16] have concatenated features from adjacent frames in making an overall articulatory classification decision, on an assumption of continuity in a local context. The *soft* method (its features denoted by A_{soft}) we argue is also appropriate because it compensates for the artifice of quantizing articulatory motion into discrete levels. Due to the bigram constraints of the recognition decoding procedure and the transcript interpolation training, motion of the recognition results within a phoneme could either signify corresponding motion of the articulators themselves, or simply an articulation "between" the rigidly assigned quantization levels. Even for the Voicing stream, which has only two classes (On or Off), a delay in voice onset time within a phoneme—especially a stop consonant—can signal a nonnative pronunciation, as investigated in [2]. Either way, for the sake of phoneme discrimination based on articulatory information, it may be important to incorporate this evidence of subphonemic

TABLE IV
PSEUDOARTICULATORY-RECOGNITION ACCURACY, IN %

	native speaker	*nonnative speakers*
jaw	52.83	65.71
lip separation	65.29	55.48
lip rounding	41.38	52.27
tongue frontness	51.54	57.24
tongue height	57.81	59.30
tongue tip	47.73	48.99
velum	76.33	44.48
voicing	81.04	69.10

motion, especially in the case of phones of interest which have expected articulatory motion within them, e.g., /t/. See Fig. 4 for a graphical depiction of this recognition-based feature extraction algorithm.

As a final note, one may be interested in the accuracy of these experiments in articulatory recognition. Such results cannot be obtained other than by artificially comparing the recognizer's output with our own interpolated articulation transcripts, and we do not really know if our articulatory expansion reflects the true articulation of our speakers, especially since they are all nonnative. This is one limitation of the data and evaluation presented here, and in the absence of a true reference we choose to use the expected mapping for the phone-level transcriptions, which we deem sufficient for the task at hand. These results are reported in Table IV, simply as a preliminary test of model suitability. They appear to be consistent with similar baseline results such as the "segment error" reported in [23], though for our nonnative Velum and Voicing streams there was an excess of insertions. Compared to the native speaker results, the better relative performance of many of the nonnative speech streams can be attributed to the larger size of the nonnative training set.

VI. EXPERIMENTS AND RESULTS

A. Experiments With Nonnative Speech

For binary classification, the Italian and German test sets consisted of equal proportions of canonical and substitution pronunciations for every target segment of interest. All canonical realizations of all segments were regarded as members of one class, and all substitution errors were regarded as members of a second class, independent of the target segment to which each error belonged. The most consistently difficult English phonemes for the German and Italian speakers in the ISLE corpus are listed in [9]—these include phonemes difficult for students to remember because of the idiosyncrasies of English orthography, and phonemes difficult for nonnative speakers to produce because they do not exist in their native language. Those used in our test sets are given in Tables II and III, along with their substitution probabilities. Our test items were taken from the ISLE corpus' Block E, F, and G recordings—read sentences typical of the kind language learners might need while traveling on a holiday or ordering food in a restaurant. Though all speakers were included in both our training and test sets,

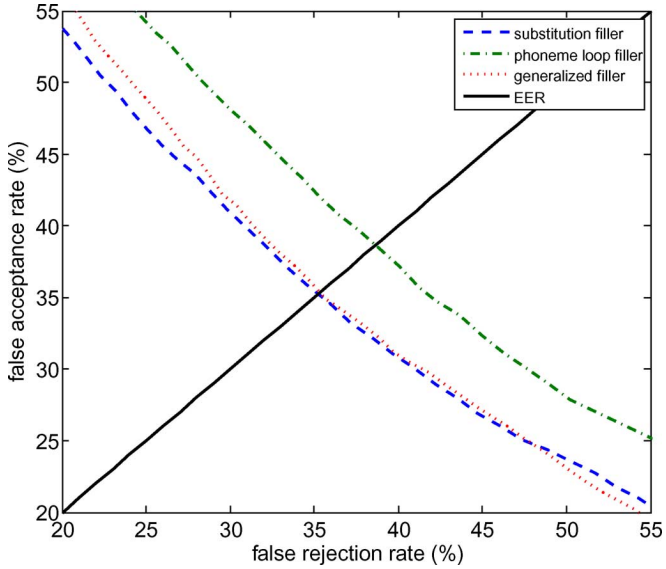

Fig. 5. DET curves for all nonnative speakers, German and Italian, over the three HMM-based confidence measures.

we argue that this is permissible given that these methods will most likely be incorporated into a language-learning system that makes use of speaker-dependent features from registered students.

Though we only examined 12 difficult target phonemes, their systematic errors account for about 30%–40% of all phone-level substitutions in the ISLE corpus. Moreover, they encompass different types of mistakes over various articulatory streams. The expected articulation for /v/ differs from its most common substitution, /f/, only in terms of Voicing. The phone /ɪ/ and its most common substitution, /i/, differ by Jaw position, Lip Separation, and Tongue Height. The Italian speakers would often insert a velar stop, /g/, after pronunciation of the velar nasal, /ŋ/, via subphonemic motion in just the Tongue Height and Velum. Our test set can be regarded as representative of the corpus at large, over many different types of substitutions.

From among the three phone-level confidence scores described in Section V-A, a baseline was selected by plotting their detection error tradeoff (DET) curves illustrating rates of false rejection versus false acceptance of the canonical pronunciation over various operating points. These curves were generated by varying the verification threshold T as explained in Section I. Results for all nonnative speakers are shown in Fig. 5. The curve with the lowest error on the equal error rate (EER) line, P_{subs}, was taken as our nonnative baseline feature.

To combine these baseline scores with the new articulatory scores proposed in Section V, we used the Weka Toolkit's implementation of the C4.5 algorithm [28]. This decision tree method struck us as the best of the available automatic techniques for several reasons. One, because it allowed for a combination of continuous and discrete features—the A_{soft} and A_{cent} articulatory recognition results alongside phonemic and articulatory confidence scores. In a study such as [27], all combinations of confidence scores were defined by optimizing a threshold on the sum of those scores, but this was not desirable

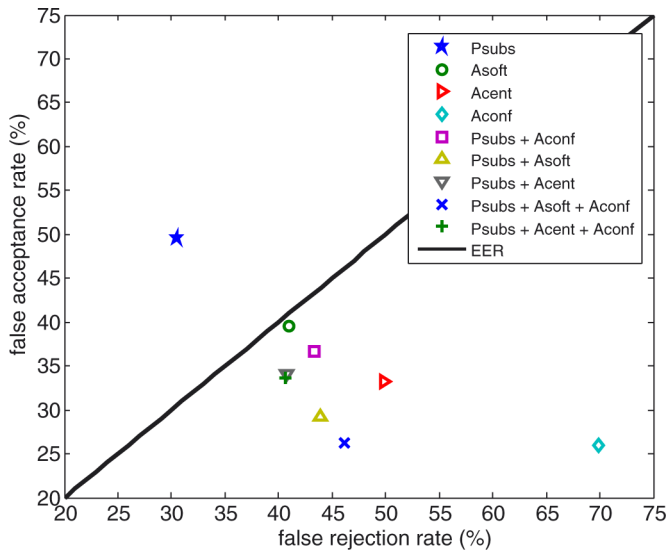

Fig. 6. German speakers' classification results over various combinations of features, for the complete test set.

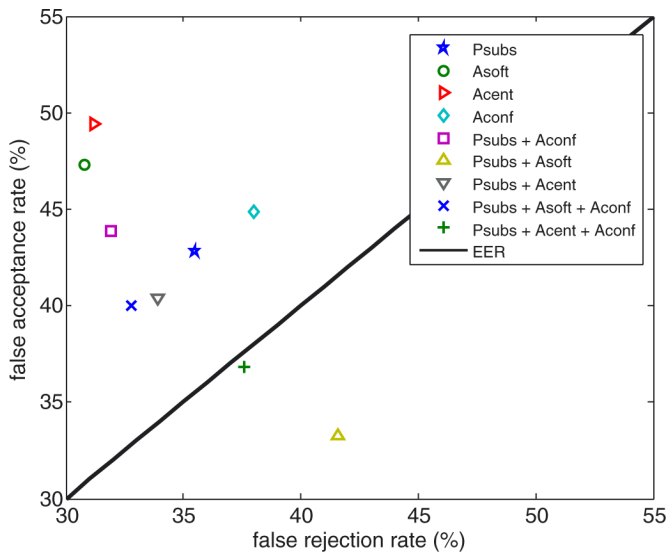

Fig. 7. Italian speakers' classification results over various combinations of features, for the complete test set.

in our case, since the A_{cent} and A_{conf} features were defined on entirely different scales. However, this did require us to report each feature combination result for just one operating point—the one determined by the trained decision tree—rather than over a range of points. The tree also facilitated the pruning of redundant features, and we expected some features to be unnecessary in certain cases. For example, all the expected substitutions for /ʌ/ when uttered by German learners of English are also vowels (/ə/, /ɒ/, /ɪ/, and /ɔː/), and so we would not expect much discrimination between canonical and erroneous pronunciations in terms of Voicing stream features. Finally, for a verification problem such as this, a decision tree's method of setting multiple sequential thresholds on many confidence-like features seemed like a logical extension of the traditional verification method of setting one threshold on just one confidence score, as introduced in Section I.

In the baseline case of just one feature—the phone-level confidence score—the tree would consist of one root node and two leaves (one each for the canonical and substitution classes), with threshold optimized by the same information gain maximization criterion used to set all thresholds in a larger tree. The trees were trained and evaluated using a tenfold cross-validation of the entire test set. Classification results in terms of false acceptance and false rejection of the canonical for various combinations of features are plotted in Figs. 6 and 7. Additional results were obtained over subsets divided by target segment, with the same test procedure, and are reported in Tables VI and VII.

B. Native Speaker Experiments

The native speaker test set, taken from the remaining half of the MOCHA-TIMIT database not used in training (one speaker), was constructed somewhat differently. Our strict definition of the canonical form in Section I does not apply to literate native speakers, who will frequently introduce nonlexical pronunciation variants into read speech—these are not really "errors" in the same sense as those made by nonnative speakers. The substitution instances for each target had to be selected artificially, based on segments which are perceptually confusable and produced in our test database, but which the speaker did not actually substitute erroneously when faced with the target prompt.

TABLE V
TARGET SEGMENTS AND THEIR SUBSTITUTIONS, FOR THE NATIVE ENGLISH SPEAKER TEST SET

target	ə	ɪ	ɛ	i	t	n	s	d	ð
substitutions	ɜː	i	ə	ɪ	p	m	ʃ	b	v
	ɛ	eɪ	eɪ	eɪ	k	ŋ	θ	g	ʒ
	əʊ	ɛ	ɜː	ɛ			f		z

Based on the most common phonemes in spoken English as reported in [29], we chose four vowels and five consonants as our targets. Substitutions for the consonants were other consonants which had the same expected manner of articulation and voicing features. Vowel substitutions were chosen based on each target's three closest vowels in Fig. 1. All targets and substitutions used in this native test set are given in Table V. As in the nonnative test sets, we maintained equal proportions of canonical and substitution instances for each target.

Fig. 8 reports the DET curves for the potential native speaker baseline scores. The confidence score derived from the substitution filler (P_{subs}) was not available in the native case because the test set was not composed of true substitutions, and so no corpus statistics were used. Numerical and graphical classification results for the native speaker are presented in Table VIII and Fig. 9; the same decision tree feature combination method as in the nonnative experiments was used here.

VII. DISCUSSION

In Fig. 5, the score with the lowest EER over the set of nonnative speech is that of the substitution filler (P_{subs}) which

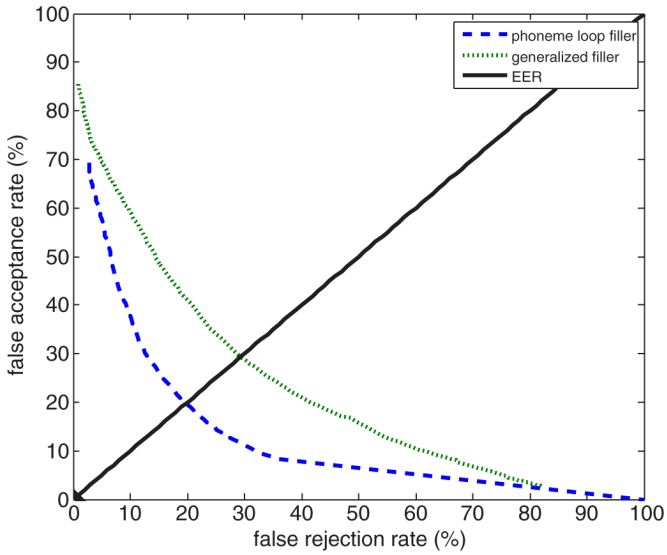

Fig. 8. DET curves for the Native British English speaker, over both HMM-based confidence measures.

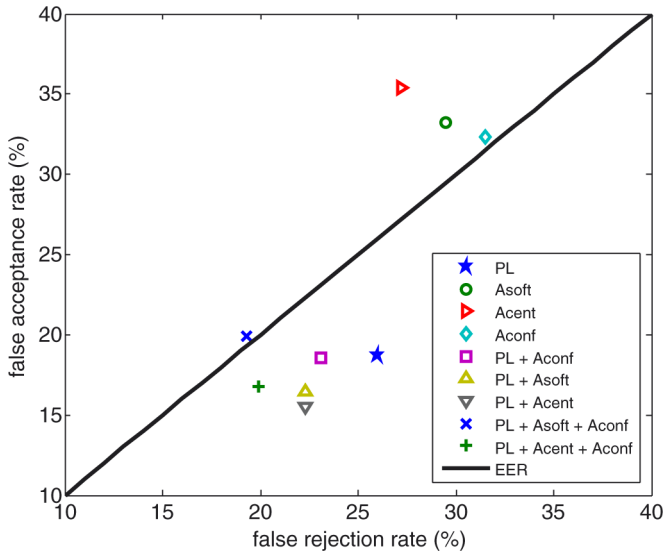

Fig. 9. Native British English speaker classification results over various combinations of features, for the complete test set.

we take as our baseline feature, though it performs only slightly better than the generalized filler, indicating that knowledge of corpus statistics is maybe not necessary to calculate an acceptable baseline confidence measure. From the two nonnative plots of various feature combinations (Figs. 6 and 7) we can discern that, in general, the new articulatory features alone (A_{conf}, A_{soft}, and A_{cent}) do not consistently perform with combined error rate lower than the baseline (which is in keeping with the findings of [17] and [27]), but when the phone-level and HAMM-derived features are combined, false acceptance is usually reduced and the combined overall error rate is lower. In foreign language practice, too many false rejections will frustrate and discourage the student, but too many false acceptances is probably less desirable, as that may undermine the learning process. Which error is more costly really depends on the student and the task, but we judged one feature combination better than another if the combined classification error rate was lower, i.e., if the operating point in Figs. 6 and 7 was closer to the origin. The baseline classification result (P_{subs}) is skewed to the side of more false acceptance than rejection, but the addition of articulatory information serves to balance the error rates, and in these plots we can see a certain reduction in the false acceptance rate for those feature sets.

The numerical nonnative classification results reported in Tables VI and VII show overall statistically significant improvements over the baseline with the addition of our HAMM-derived features, but this is not consistent for every target segment subset. Here a significant reduction in the combined error rate is defined by a p-value of 0.05 or less using McNemar's test (denoted in bold in the tables). Even in the case of an overall improvement, often a large reduction in false rejection is accompanied by a small increase in false acceptance, or vice versa. If some of the target-dependent subsets allowed for less (or no) improvement over the baseline, this is probably because the baseline score generalizes well to discriminate between the canonical form and its substitutions in such a limited test set. Over the complete test set, just one baseline threshold may not be optimal for all target words, and so the addition of articulatory features offers significant reductions in classification error.

The best feature set overall, for both Italian and German speakers, was the $P_{\text{subs}} + A_{\text{soft}} + A_{\text{conf}}$ combination, which offered a 3%–4% absolute reduction in combined classification error rate over the baseline for the complete test set, and as high as a 16%–17% absolute improvement for individual segments (e.g., /t/ for the German speakers). However, this combination did not perform significantly better than $P_{\text{subs}} + A_{\text{cent}} + A_{\text{conf}}$, so we cannot conclude that the *soft* method of extracting articulatory recognition-based features was more appropriate than the *centered* method.

More than the Germans, the Italian speakers had several individual segments with error rates not reduced by the addition of articulatory information. This difference in results for the two speaker types can be attributed to the fact that the Italian speakers were generally less proficient in English than the German speakers [9]. We can infer that the Italians' mistakes in pronunciation were more dramatic and obvious than those of the Germans who, with higher proficiency and a native language more closely related to English, probably introduced more subtle substitutions. Classification of these minor differences in articulation stands to benefit the most from articulatory features which bear pronunciation details not captured by phone-level HMMs.

If our nonnative results seem modest in light of the fact that these were all binary classifications tested on equal proportions of each class, consider a couple of things. First, in [17] the overall absolute improvement in basic word recognition offered by combining traditional and articulatory models similar to these was on the order of 1%–2%. Second, that [9] reported an interannotator agreement of "at best" 70% when simply

TABLE VI
RESULTS FOR GERMAN SPEAKERS, REPORTED AS FALSE REJECTION RATE/FALSE ACCEPTANCE RATE (IN %). ENTRIES IN BOLD WERE SIGNIFICANTLY BETTER THAN THE BASELINE (P_{subs}) WITH $p \leq 0.05$ USING MCNEMAR'S TEST

target	*instances*	P_{subs}	$P_{subs}+$ A_{conf}	$P_{subs}+$ A_{soft}	$P_{subs}+$ A_{cent}	$P_{subs}+$ $A_{soft}+A_{conf}$	$P_{subs}+$ $A_{cent}+A_{conf}$
z	596	62.1 / 12.1	61.4 / 15.4	40.3 / 33.6	47.3 / 26.2	38.3 / 36.9	43.6 / 30.9
ə	1532	8.9 / 77.3	32.6 / 50.7	38.0 / 60.3	**28.9 / 51.2**	**40.1 / 38.5**	42.2 / 38.5
v	310	64.5 / 13.5	39.4 / 30.3	**29.7 / 27.7**	**37.4 / 22.6**	**36.1 / 29.0**	**39.4 / 22.6**
ʌ	522	42.5 / 18.8	33.3 / 23.4	**33.0 / 18.0**	32.6 / 26.1	34.5 / 22.6	32.6 / 28.0
t	344	20.3 / 56.4	**23.8 / 23.3**	**27.9 / 14.5**	**37.8 / 23.3**	**22.7 / 20.9**	**17.4 / 26.2**
all	3304	30.5 / 49.6	43.3 / 36.6	**43.9 / 29.3**	**40.7 / 34.1**	**46.1 / 26.2**	**40.6 / 33.7**

TABLE VII
RESULTS FOR ITALIAN SPEAKERS, REPORTED AS FALSE REJECTION RATE/FALSE ACCEPTANCE RATE (IN %). ENTRIES IN BOLD WERE SIGNIFICANTLY BETTER THAN THE BASELINE (P_{subs}) WITH $p \leq 0.05$ USING MCNEMAR'S TEST

target	*instances*	P_{subs}	$P_{subs}+$ A_{conf}	$P_{subs}+$ A_{soft}	$P_{subs}+$ A_{cent}	$P_{subs}+$ $A_{soft}+A_{conf}$	$P_{subs}+$ $A_{cent}+A_{conf}$
ʊ	312	0.6 / 87.8	59.0 / 22.4	43.6 / 47.4	32.1 / 50.6	67.9 / 12.8	39.7 / 48.1
ɪ	1946	44.2 / 34.9	**29.3 / 42.4**	**24.8 / 44.6**	**25.1 / 45.6**	**22.2 / 48.8**	30.5 / 43.7
ʌ	918	31.4 / 31.4	34.2 / 25.1	29.6 / 32.7	26.4 / 34.6	31.2 / 32.7	29.6 / 34.2
ə	2500	32.1 / 54.9	**32.6 / 51.2**	33.5 / 50.2	32.6 / 51.4	31.4 / 55.2	**37.4 / 43.9**
t	1388	20.6 / 53.9	**20.3 / 29.5**	**30.5 / 34.9**	30.0 / 41.2	**23.3 / 25.2**	**19.7 / 32.3**
ŋ	238	31.9 / 36.1	21.8 / 50.4	37.0 / 32.8	32.8 / 35.3	32.8 / 42.9	32.8 / 46.2
ɜː	566	41.7 / 30.4	36.4 / 38.2	48.1 / 27.6	45.6 / 24.0	37.1 / 32.5	37.5 / 35.7
all	7868	35.5 / 42.8	**31.9 / 43.9**	**41.6 / 33.3**	**33.9 / 40.4**	**32.8 / 40.0**	**37.6 / 36.8**

TABLE VIII
NATIVE BRITISH ENGLISH SPEAKER RESULTS, REPORTED AS FALSE REJECTION RATE/FALSE ACCEPTANCE RATE (IN %). ENTRIES IN BOLD WERE SIGNIFICANTLY BETTER THAN THE BASELINE (PL) WITH $p \leq 0.05$ USING MCNEMAR'S TEST

target	*instances*	PL	$PL+$ A_{conf}	$PL+$ A_{soft}	$PL+$ A_{cent}	$PL+$ $A_{soft}+A_{conf}$	$PL+$ $A_{cent}+A_{conf}$
all	4896	26.0 / 18.7	**23.1 / 18.6**	**22.3 / 16.5**	**22.3 / 15.5**	**19.3 / 19.9**	**19.9 / 16.8**

deciding the phone-level location of a pronunciation error (but not deciding what the substitution is), which is in essence what our automatic classifier did. This indicates that often it is not a trivial decision to attribute an abstract substitution to a single phone location, and that perhaps pronunciation error annotation on the subphone or suprasegmental level would achieve higher interlabeler agreement.

We found similar results for the supplementary native English speaker experiments (Fig. 9 and Table VIII), though with slightly lower error rates overall, probably because the "substitution" instances chosen artificially were easier to discriminate from their so-called targets because they were not true pronunciation mistakes but phonemes known to be perceptually confusable. For example, while the artificial errors for consonants shared the target's voicing and manner, real errors for consonants were in many cases pronounced both with the same manner and place of articulation as the target (see Tables II and III), therefore making real errors more difficult to classify automatically. In the native speaker case, the score derived from the phoneme loop filler (PL) was our obvious choice for a baseline feature, from Fig. 8. Just as for the nonnative speakers, the addition of articulatory features offered significant improvements over the baseline error rates, both for false acceptance and rejection. This suggests the usefulness of our proposed features for utterance verification outside the domain of second-language pedagogy.

VIII. CONCLUSION

We have demonstrated a useful and unique method of assigning an articulatory representation to a phone-level transcription. Hidden articulator Markov models trained on this representation were used to generate novel articulatory confidence measures and recognition-based feature vectors. These vectors successfully reduced error rates in segment-level pronunciation classification when combined with traditional segmental confidence scores, for two types of nonnative speakers and over a native speaker test set as well. It is remarkable to consider this improvement in light of the fact that the articulatory models had the same topology as the phoneme models, were trained on the same spectral features as the phoneme models, needed no direct articulatory measurement or transcription, and required no

TABLE IX
EXPECTED BRITISH ENGLISH VOWEL ARTICULATIONS

IPA phone	*ISLE phone*	*example word*	*jaw*	*lip separation*	*lip rounding*	*tongue frontness*	*tongue height*	*tongue tip*	*velum*	*voicing*
ɑː	AA	**bard**	3	2	1	1	0	0	0	1
æ	AE	**bad**	3	3	2	3	0	0	0	1
ʌ	AH	**bud**	2	2	2	2	0	0	0	1
ɔː	AO	**bawd**	3	2	1	0	2	0	0	1
aʊ	AW	**bowed**	3	2	2	2	0	0	0	1
			1	2	0	1	2	0	0	1
ə	AX	**about**	2	2	2	2	1	0	0	1
aɪ	AY	**bide**	3	2	2	2	0	0	0	1
			1	2	3	3	2	0	0	1
ɛ	EH	**bed**	3	2	2	3	1	0	0	1
ɜː	ER	**bird**	2	2	2	2	1	0	0	1
eɪ	EY	**bayed**	1	2	3	4	2	0	0	1
ɪ	IH	**bid**	3	2	3	4	2	0	0	1
i	IY	**bead**	0	1	3	4	3	0	0	1
ɒ	OH	**body**	2	1	2	0	1	0	0	1
əʊ	OW	**bode**	3	2	1	3	2	0	0	1
			2	1	0	1	2	0	0	1
ɔɪ	OY	**boy**	2	2	1	0	1	0	0	1
			1	2	3	3	2	0	0	1
ʊ	UH	**buddhist**	1	2	1	1	2	0	0	1
uː	UW	**booed**	1	1	0	1	3	0	0	1

prior knowledge of the corpus substitution statistics. This augmented representation helped us to model subtler differences in pronunciation by factoring each segment into its relevant discriminative components.

IX. FUTURE DIRECTIONS

This method of locating a segment-level pronunciation mistake represents one important step in the complex pronunciation evaluation task. Pronunciation results generated by this algorithm should in the future be used in combination with other—perhaps suprasegmental—scores to derive an overall pronunciation score for a given speaker or utterance. The feature sets could benefit from prosodic information when, for example, distinguishing between full and reduced vowels, or stressed and unstressed syllables.

Our proposed transcription expansion method is by no means the only mapping that improves close phonemic discrimination over traditional models—there is still room for optimization, depending on the task. Though we incorporated a large amount of prior information about expected English substitutions based on our speaker set, future researchers in this area might also include phonological constraints from the speakers' native language—and not just the language they are learning—when expanding and interpolating the articulatory transcriptions. For example, what allophonic variants are native German and Italian speakers likely to produce when reading English? In addition to segment-level substitutions, what articulation should we expect based on Italian or German phonology? These are relevant open questions.

Future work in this area may also choose to rely on direct articulatory measurement for model training purposes. The models would better reflect the true articulation, rather than just the expected form. Such models could potentially have a finer-grain resolution than those set forth in Table I. Both these improvements may serve to make the discrimination between canonical and erroneous segments more accurate. No existing collection of articulatory data has the same degree of nonnative speaker variability as seen in the ISLE corpus, nor the same focus on second language acquisition and pedagogy, and until such a corpus is created, the use of direct articulatory measurement in this domain will be limited.

Finally, though we have shown that an assumption of dependence among the streams in our representation is not necessary to effect an improvement in verification accuracy, it may be worthwhile to try modeling such a dependency in, for example, a dynamic Bayesian network framework. As explained in Section III, this approach would probably make articulatory recognition more accurate, but may also decrease the variation in recognition results—variation that may help signify the presence of a pronunciation mistake. Such a network could be implemented without any changes to our *centered* or *soft* methods of calculating features from the articulatory recognition results.

APPENDIX

Tables IX and X display the expected articulatory mappings for British English vowels and consonants, respectively, derived from [1] and [17]. For those with subphonemic motion, only

TABLE X
EXPECTED BRITISH ENGLISH CONSONANT ARTICULATIONS

IPA phone	*ISLE phone*	*example word*	*jaw*	*lip separation*	*lip rounding*	*tongue frontness*	*tongue height*	*tongue tip*	*velum*	*voicing*
b	B	**b**et	1	0	2	2	1	1	0	1
			1	2	2	2	1	1	0	1
d	D	**d**ebt	1	1	2	4	3	4	0	1
			1	2	2	4	2	3	0	1
ɡ	G	**g**et	1	2	2	0	3	1	0	1
			1	2	2	0	2	1	0	1
p	P	**p**et	1	0	2	2	1	1	0	0
			1	2	2	2	1	1	0	0
t	T	**t**at	1	1	2	4	3	4	0	0
			1	2	2	4	2	3	0	0
k	K	**c**at	1	2	2	0	3	1	0	0
			1	2	2	0	2	1	0	0
ð	DH	**th**at	2	2	2	4	2	2	0	1
θ	TH	**th**in	2	2	2	4	2	2	0	0
v	V	**v**an	2	0	2	2	1	1	0	1
f	F	**f**an	2	0	2	2	1	1	0	0
z	Z	**z**oo	1	2	2	3	3	3	0	1
s	S	**s**ue	1	2	2	3	3	3	0	0
ʒ	ZH	measure	2	2	1	3	3	0	0	1
ʃ	SH	**sh**oe	2	2	1	3	3	0	0	0
ʤ	JH	**j**eep	2	2	2	4	3	4	0	1
			1	2	1	3	3	0	0	1
ʧ	CH	**ch**eap	2	2	2	4	3	4	0	0
			1	2	1	3	3	0	0	0
m	M	**m**et	1	0	2	2	1	1	1	1
n	N	**n**et	1	1	2	2	3	4	1	1
ŋ	NG	thi**ng**	1	2	2	0	3	1	1	1
l	L	**l**ed	1	2	2	3	2	4	0	1
ɹ	R	**r**ed	1	2	1	2	2	3	0	1
w	W	**w**ed	1	2	0	0	3	1	0	1
j	Y	**y**et	1	2	2	4	3	3	0	1
h	HH	**h**at	2	2	2	2	1	1	0	0

the start and end positions are shown here. See Table I for the articulatory classes represented by these integers.

REFERENCES

[1] P. Ladefoged, *A Course in Phonetics*, 5th ed. Boston, MA: Thomson Wadsworth, 2006.

[2] A. Kazemzadeh, J. Tepperman, J. Silva, H. You, S. Lee, A. Alwan, and S. Narayanan, "Automatic detection of voice onset time contrasts for use in pronunciation assessment," in *Proc. InterSpeech ICSLP*, Pittsburgh, PA, 2006, paper 1884-Mon3FoP.8.

[3] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. ICASSP*, Philadelphia, PA, 2005, pp. 937–940.

[4] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *J. Acoust. Soc. Amer.*, vol. 111, no. 2, pp. 1086–1101, 2002.

[5] K. Livescu and J. Glass, "Feature-based pronunciation modeling for speech recognition," in *Proc. HLT/NAACL*, Boston, MA, 2004.

[6] H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan, "Pronunciation variations of Spanish-accented English spoken by young children," in *Proc. Eurospeech*, Lisbon, Portugal, 2005, pp. 749–752.

[7] N. Mote, A. Sethy, J. Silva, S. Narayanan, and L. Johnson, "Detection and modeling of learner speech errors: The case of Arabic tactical language training for American English speakers," in *Proc. InStil*, Venice, Italy, 2004, paper 012.

[8] R. Delmonte, M. Petrea, and C. Bacalu, "SLIM prosodic module for learning activities in a foreign language," in *Proc. ESCA, Eurospeech'97*, Rhodes, Greece, 1996, vol. 2, pp. 669–672.

[9] E. Atwell, P. Howarth, and C. Souter, "The ISLE corpus: Italian and German spoken learner's English," *ICAME J.*, vol. 27, pp. 5–18, 2003.

[10] J. Caminero, C. de la Torre, L. Villarrubia, C. Matin, and L. Hernandez, "On-line garbage modeling with discriminant analysis for utterance verification," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 2111–2114.

[11] J. Dolfing and A. Wendemuth, "Combination of confidence measures in isolated word recognition," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 3237–3240.

[12] P. Ramesh, C.-H. Lee, and B.-H. Juang, "Context dependent anti subword modeling for utterance verification," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 3233–3236.

[13] D. Willett, A. Worm, C. Neukirchen, and G. Rigoll, "Confidence measures for HMM-based speech recognition," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 3241–3244.

[14] J. P. Campbell , Jr., "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[15] K. Leung, M. Mak, and S. Kung, "Applying articulatory features to telephone-based speaker verification," in *Proc. ICASSP'04*, Montreal, QC, Canada, 2004, pp. I-85–I-88.
[16] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, Univ. Bielefeld, Bielefeld, Germany, 1999.
[17] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models for speech recognition," *Speech Commun.*, vol. 41, no. 2, pp. 511–529, Oct. 2003.
[18] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *Proc. ICASSP'05*, Philadelphia, PA, 2005, pp. 1009–1012.
[19] J. Tepperman and S. Narayanan, "Hidden-articulator Markov models for pronunciation evaluation," in *Proc. ASRU*, San Juan, Puerto Rico, 2005, pp. 174–179.
[20] A. Wrench, "The MOCHA-TIMIT Articulatory Database," [Online]. Available: http://www.cstr.ed.ac.uk/artic/mocha.html
[21] E. Grabe, "Intonational variation in urban dialects of English spoken in the British Isles," in *Regional Variation in Intonation*, ser. Linguistiche Arbeiten, P. Gilles and J. Peters, Eds. Tuebingen, Germany: Niemeyer, 2004, pp. 9–31.
[22] A. Gutkin and S. King, "Detection of symbolic gestural events in articulatory data for use in structural representations of continuous speech," in *Proc. ICASSP'05*, Philadelphia, PA, 2005, pp. I-885–I-888.
[23] K. Hacioglu, B. Pellom, and W. Ward, "Parsing speech into articulatory events," in *Proc. ICASSP'04*, Montreal, QC, Canada, 2004, pp. I-925–I-928.
[24] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models: Performance improvements and robustness to noise," in *Proc. ICSLP*, Beijing, China, 2000, pp. 131–134.
[25] S. Young *et al., The HTK Book*. Cambridge, U.K.: Univ. Cambridge, 2002 [Online]. Available: http://htk.eng.cam.ac.uk/
[26] M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic bayesian networks," in *Proc. IEICI Beyond HMM Workshop*, Kyoto, Japan, 2004, pp. 37–42.
[27] K.-Y. Leung and M. Siu, "Articulatory-feature-based confidence measures," *Comput. Speech Lang.*, vol. 20, pp. 542–562, 2006.
[28] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.
[29] P. Denes, "On the statistics of spoken English," *J. Acoust. Soc. Amer.*, vol. 35, no. 6, pp. 892–904, Jun. 1963.

Joseph Tepperman (S'07) received the B.S. and M.S. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, in 1999 and 2003, respectively. He is currently pursuing the Ph.D. degree at USC's Signal Analysis and Interpretation Laboratory (SAIL). The focus of his research is pronunciation modeling and evaluation, using articulatory and prosodic cues.

Actively involved with the Technology-Based Assessment of Language and Literacy (Tball) project in collaboration with UCLA, he is also interested in speech recognition applications for art and music.

Shrikanth S. Narayanan (S'88–M'95–SM'02) received the Ph.D. degree from the University of California, Los Angeles, in 1995.

He is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, where he holds appointments as Professor in electrical engineering and jointly in computer science, linguistics, and psychology. Prior to joining USC, he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal Member of its Technical Staff from 1995–2000. At USC, he is a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, an NSF Engineering Research Center. He has published over 235 papers and has 14 granted/pending U.S. patents.

Dr. Narayanan is a recipient of an NSF CAREER Award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost Fellowship from the USC Center for Interdisciplinary Research, a Mellon Award for Excellence in Mentoring, and a recipient of a 2005 Best Paper Award from the IEEE Signal Processing Society. Papers by his students have won best student paper awards at ICSLP'02, ICASSP'05, and MMSP'06. He is an Editor for the *Computer Speech and Language Journal* (2007-present) and an Associate Editor for the IEEE *Signal Processing Magazine*. He was also an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2004). He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society and the Speech Communication Committee of the Acoustical Society of America. He is a Fellow of the Acoustical Society of America and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu.