Speech production knowledge in automatic speech recognition

Speech production knowledge in automatic speech recognition

Simon King[a] and Joe Frankel
Centre for Speech Technology Research, University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW United Kingdom

Karen Livescu
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Room 32-G482
Cambridge MA 02139 USA

Erik McDermott
Nippon Telegraph and Telephone Corporation, NTT Communication Science Laboratories
2–4 Hikari-dai, Seika-cho, Soraku-gun
Kyoto-fu 619-0237 Japan

Korin Richmond and Mirjam Wester
Centre for Speech Technology Research, University of Edinburgh

(Dated: October 31, 2006)

Although much is known about how speech is produced, and research into speech production has resulted in measured articulatory data, feature systems of different kinds and numerous models, speech production knowledge is almost totally ignored in current mainstream approaches to automatic speech recognition. Representations of speech production allow simple explanations for many phenomena observed in speech which cannot be easily analyzed from either acoustic signal or phonetic transcription alone. In this article, we provide a survey of a growing body of work in which such representations are used to improve automatic speech recognition.

PACS numbers: 43.72.Ne (Automatic speech recognition systems); 43.70.Jt (Instrumentation and methodology for speech production research); 43.70.Bk (Models and theories of speech production)
Keywords: Automatic speech recognition; Speech production; Articulatory features; Articulatory inversion

I. INTRODUCTION

There is a well-established body of knowledge about the speech production mechanism (e.g., Löfqvist, 1997), covering articulatory processes (e.g., Perkell, 1997), coarticulation (e.g., Farnetani, 1997) and so on. The movements of the articulators can be directly measured in real time and models of the articulatory system, particularly the tongue, have been built (e.g., Honda *et al.*, 1994; Kaburagi and Honda, 1996). Aside from actual articulatory data, other representations of speech production are available, including various kinds of feature systems, gestures and landmarks. These expressive representations allow simple explanations for many phenomena observed in speech which cannot be easily analyzed from either the acoustic signal or the phonetic transcription alone.

Acoustic modeling for automatic speech recognition (ASR) currently uses very little of this knowledge and as a consequence speech is only modeled as a *surface* phenomenon. Generally, hidden Markov models (HMMs) link the phonetic level to the observed acoustic signal via a single discrete hidden variable, the state. The state space (i.e., the set of values the state variable can take, which may be many thousands) is a homogeneous layer with no explicit model of the structural differences or similarities between phones; the evolution of the state through time is modeled crudely. In this article we consider whether acoustic modeling for speech recognition stands to benefit from the judicious use of knowledge about speech production.

The standard approach to acoustic modeling continues to be the "beads on a string" model (Ostendorf, 1999) in which the speech signal is represented as a concatenation of phones. The fact that the acoustic realization of phones is context-dependent – the consequence of coordinated motion of multiple, slow-moving physical articulators – is accounted for by the use of context-dependent models. Because the set of contexts is very large, statistical clustering techniques must be used. This approach is probably not optimal, and perhaps effective only for relatively constrained types of speech. Although these clustering techniques use articulatory/phonetic features, a direct use of these features as statistical factors may offer better performance. Variations in speech production

[a] Electronic address: Simon.King@ed.ac.uk

S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester. Speech production knowledge in automatic speech recognition. Journal of the Acoustical Society of America, 121(2):723-742, February 2007.

(e.g., due to speaking rate) are either not modeled, or require the creation of situation-specific models, leading to a problem of robust estimation. There is some work on explicit modeling of pronunciation variation, but this is severely limited by the coarseness of the phone unit: another consequence of the "beads on a string" approach. Section II provides a full definition of what we mean by "speech production knowledge". For now, we can take it to include data about articulation (recorded directly, or annotated by human labelers), automatic recovery of such data, or features derived from phonetic transcriptions.

A. The scope of this article

We aim to provide a comprehensive overview of the large body of speech recognition research that uses speech production knowledge: a road map for the reader that makes connections between the different approaches. This article was inspired by the Beyond HMM Workshop (2004) and had its genesis in a short paper by McDermott (2004).

Section I B makes the case for using speech production knowledge, Section I C acknowledges ways in which current systems already do and Section I D gives some pointers to early work. Section II defines in detail just what is meant by "speech production knowledge" and discusses how it can be represented with a view to using it in speech recognition; we do not attempt a complete review of all work on speech production, nor do we consider theories of speech *perception* with a production basis (Liberman and Mattingly, 1985). We will not consider formant frequencies in any depth (because they only give an incomplete picture of speech production) or prosody, source features or phrasal effects (because we are not aware of any production-based ASR system that uses them). In Sections III and IV, we look at how this speech production knowledge can be obtained, whether it is from articulatory measurements, manual transcription, derivations from phonetic labels or by machine-learning. Section V provides a survey of efforts to use production knowledge in acoustic modeling for automatic speech recognition. Finally, in Section VI, we highlight some ongoing work and suggest future directions.

B. The case for using speech production knowledge in speech recognition

Incorporating speech production knowledge into ASR may alleviate some of the problems outlined earlier and enable improved recognition of spontaneous speech, greater robustness to noise, and multi-lingual acoustic models (Rose *et al.*, 1996). In practice, it is hard to realize these benefits.

Most production representations use a **factored representation**: parallel "streams" of features/parameters. Since any given feature/parameter will typically be shared amongst many phoneme classes, the training data are used in a potentially more effective way. Low-frequency phonemes will share features with high-frequency ones, benefiting from their plentiful training data. Parameter tying for context-dependent HMMs already takes advantage of this property.

Rather than modeling complex *acoustic* effects of co-articulation, **explicit modeling** at the production level specifies precisely where, when and how co-articulation occurs. Since production representations are easily interpreted, models that use them are more **transparent** than HMMs, where the hidden state defies any interpretation or post-hoc error analysis. Although our understanding of the modeling capabilities of HMMs has recently been advanced (Bilmes, 2004; Bridle, 2004; Tokuda *et al.*, 2004), there is still a long way to go before we are able to interpret current large systems.

The advantages of explicit modeling and a factored representation together imply better performance on **spontaneous or casual speech** because of the greater degree of co-articulation observed in this type of speech (Farnetani, 1997). We also expect production-based models to be **noise robust**. The factored representation means each feature/parameter is easier to recognize than, say, 61 phoneme classes, because features/parameters typically have far fewer than 61 possible values. In a factored representation, errors for each factor are multiplied together, which could potentially make the situation worse, but each feature/parameter will be affected differently by noise, so we could expect that – provided there is a little redundancy, or a strong enough language model – in the presence of noise, sufficient features/parameters could still be identified ("islands of reliability") to perform speech recognition. In order to take full advantage of the varying reliability of the different features, a confidence measure is required.

It is possible to construct production-based representations that are **multilingual** or perhaps even **language universal**. This is an under-explored area, deserving of further research. The International Phonetic Alphabet (IPA, International Phonetic Association (1999)) provides a phoneme set which is intended to be universal, but suffers from a number of problems, such as: A single IPA symbol may be pronounced somewhat differently in different languages; Some symbols are very rare amongst the world's languages. Features offer a powerful and language-universal system (Ladefoged, 1997). Some representations of speech production can be argued to be relatively **speaker independent**, compared to standard spectral features (Maddieson, 1997). Expressing **pronunciation variation** as phonemic transcriptions is problematic (Section V E). A factored feature representation is potentially both more expressive (e.g., it allows small variations that do not result in phonemes changing class) and more compact.

C. How much production knowledge do current HMM systems use?

Current systems, which are almost invariably HMM-based, use a little knowledge of speech production. One of the above advantages of a speech production representation – its factorial nature – is exploited, albeit to a limited extent and in somewhat opaque fashion, by standard HMM systems during decision tree-based parameter tying. Vocal tract length normalization (Cohen *et al.*, 1995) acknowledges a physiological fact of speech production, and is widely used. Jurafsky *et al.* (2001) suggest that models of triphones (context-dependent phones) can usually deal with phonetic substitutions, but not large-scale deletions.

1. Decision trees for state tying

Because the state space of a context-dependent HMM system is large (there are more parameters than can be learned from current data sets), it is necessary to share parameters within clusters of states. Bundles of discrete features, similar to the systems discussed in Section II B, are the usual representation used for phonetic context when building the decision trees used for tying the states of context-dependent HMMs such as triphone models (e.g., Young *et al.*, 2002). If phonemes were used to describe the left and right context, the power of state-tying would be severely restricted. For example, it would not be possible to identify the similar effect that nasals like [n] and [m] have on preceding vowels, or the similar formant transitions seen in vowels following bilabial stops like [p] and [b]. The fact that [p] and [b] have the same values for some discrete features tells us that they are similar (in terms of production) and will have similar effects on neighboring phones. This use of features is still limited, because features are attached to phones (or HMM states), so the power of the modeling is still restricted by the "beads on a string" problem. To really exploit the power of features to describe contextual and pronunciation variation in speech, probably requires the model to retain an internal feature-based representation.

2. State tying gives structure to the state space

After state tying, the state space has structure. When a pair of states from different triphone models (the same base phone in different left and/or right contexts) are tied, the acoustics of these two triphones must be similar. Since the tying was guided by features that can be related to production knowledge, the underlying production of the two triphones must also be similar. A cluster of tied states thus connects a localized region of acoustic space with a localized region of articulatory space. Therefore, although the state-space structure (the set of clusters) formed by state tying is not at all easy to interpret, it forms a mapping from production to acoustics. Within a cluster, the mapping is approximately constant (all states in the cluster share the same values for some subset of the production-based features and all use the same output density); from cluster to cluster, the mapping changes, sometimes by a small amount (a smooth region in the global articulation-to-acoustic mapping), sometimes by a large amount (a discontinuity in the global articulation-to-acoustic mapping). However, little attempt is made to take any further advantage of this implicit mapping, by interpolating between clusters for example. Luo and Jelinek (1998) suggested "nonreciprocal data sharing" as a method for estimating HMM state parameters without hard state tying. This is, in essence, an interpolation between the clusters of states, but it does not explicitly use any articulatory information.

D. Historical perspective

Early attempts to use knowledge of speech production for speech recognition were limited. Since articulatory motion data was not easily available, knowledge had to be gleaned from human examination of the acoustic signal, from experiments on human reading of spectrograms (Cole *et al.*, 1980; Zue, 1985), from static X-ray images or introspection. This knowledge was then translated into rules or other classifiers that typically produced either a phoneme string or a phoneme lattice as output. Some highlights of this early work are mentioned below.

Fujimura (1986) proposed that certain, critical articulatory movements are more speaker-invariant than the acoustic signal. Cole *et al.* (1986) suggested that features (e.g., formant-related) were required to make fine phonetic distinctions and De Mori *et al.* (1976) used features attached to syllables. Lochschmidt (1982) used simple articulatory parameters to aid phonetic classification. Several systems have used a variety of acoustic-phonetic features, which often have some speech production basis. The CMU Hearsay-II system (Goldberg and Reddy, 1976) and the CSTR Alvey recognizer (Harrington, 1987) made use of phonetic features. More recent work has continued this knowledge-based approach (Bitar and Espy-Wilson, 1995, 1996; Espy-Wilson and Bitar, 1995).

In this article, we will not further discuss these early (and often failed) attempts, which generally used knowledge-based "expert systems" methods. We will instead consider more recent work, which uses statistical/machine-learning techniques. One of the earliest attempts at a production-inspired model within a statistical machine-learning framework, was the Trended HMM of Deng (1992) in which HMM states, instead of generating independent identically distributed observations, generate "trended" sequences of observations in the observation space. The model only accounts for one aspect of the acoustic consequences of speech production: piecewise smooth/continuous observation sequences. It does not attempt to explicitly model underlying production.

II. WHAT IS "SPEECH PRODUCTION KNOWLEDGE"?

In this section, we examine what various researchers mean by "speech production knowledge", the linguistic theories which provide the original motivations, and how speech production can be represented, whether that is in a literal or abstract form.

A. Why the representation matters

For production knowledge to be useful for speech modeling and recognition, an encoding must be chosen. A variety of possibilities exist, ranging from continuous-valued measurements of vocal tract shape during speech, to the use of discrete-valued manually selected feature sets. The speech production parameters which appear in this survey can be broadly categorized as discrete or continuous. "Discrete" is used in this article to cover both categorical features and discretized positions (e.g., high/mid/low are the three possible values for the discrete feature "height" in some feature systems). As we will see, the form of the representation is crucial. If we adopt an approach that models the representation explicitly, the representation will directly determine the type of statistical or other model that can be used.

Generally, representations abstract away from the articulatory organs and lie somewhere between a concrete description of the continuous-valued physical positions and motions of the articulators and some higher-level symbolic, linguistic representation (e.g., phonemes). The motivations for each representation are quite different: a desire to explain co-articulation, or the atomic units in some particular phonological theory, for example. Likewise, the position of each representation along the *physical production space* $\longleftrightarrow$ *abstract linguistic space* axis is different. All of them claim to normalize articulation, within and particularly across speakers – they are more speaker-independent than measurements of tongue position, for example. Most also claim to be language independent (Ladefoged, 1997). Many claim to be able to explain phenomena that have complex acoustic consequences, such as co-articulation or phonological assimilation, quite simply, e.g., by overlapping or spreading of features (Farnetani, 1997). These are all strong motivations for believing that production knowledge should be used in ASR.

Clearly, the degree of abstraction affects the usefulness of a representation for modeling purposes. Whilst physical measurements of articulator positions might be most true to the reality of speech production, they pose significant problems for statistical modeling – for example, they are generally continuous-over-time trajectories and therefore require a different class of models and algorithms than frame-based data. At the other extreme, highly abstract representations might be simpler to model, but cannot express the details of speech production that might improve speech recognition accuracy. Typically, more abstract representations will tend to be discrete whereas concrete ones will tend to be continuous-valued. Discretization of continuous processes is common when formulating numerical models. In the case of speech production, the choice of symbols may be suggested by the feature system in use (e.g., a traditional place/manner system may have 9 values for place and 5 for manner) or by quantizing articulatory measurements (e.g., Stephenson's work discussed in Section V A 3). We will use the terms *articulatory features (AFs)* to refer to discrete-valued representations and *articulatory parameters* to refer to continuous-valued representations.

B. Discrete representations of speech production

Discrete representations of speech production fall into two categories. In one, the number of features is usually small, with each feature taking a value from a set of possible values. It is possible for a feature to have an unspecified value (the set of features is then called "underspecified"). These features are often associated with a linguistic unit. A traditional system for describing phonemes using a small number of features, each of which can take multiple values, has as its two most important features manner and place. To these, various other features can be added; for example, the inventory used by Kirchhoff (1999) is manner (possible values: vowel, lateral, nasal, fricative, approximant, silence), place (dental, coronal, labial, retroflex, velar, glottal, high, mid, low, silence), voicing (voiced, voiceless, silence), rounding (rounded, unrounded, nil, silence) and front-back (front, back, nil, silence). Because the set of possible places of articulation depends on manner, the values that place can take may be made conditional on the value of manner (Chang *et al.*, 2005; Juneja and Espy-Wilson, 2003b). This is a frequently-used representation for modeling, where the features are known as *pseudo-articulatory* features, or simply articulatory features. AF-labeled data are commonly produced using rule-based systems which map from existing labels to a corresponding articulatory configuration or sequence of configurations (see Section IV B 3). Other discrete parameterizations include quantizing measured articulatory data (Section V A 3).

The other category of representations uses a larger number of *binary* features; a vector of such features may be associated with a linguistic unit or, for the purposes of ASR, may be specified for every time frame. One influential phonological model (Chomsky and Halle, 1968) represents phonemes as vectors of binary features, such as voiced/voiceless, nasal/non-nasal or rounded/unrounded. These all have a physical production interpretation, although they were intended for use in phonological rules. Some approaches to ASR described in this article use this approach (Section V A 1). However, they generally adopt only the feature set and ignore the rule-based phonological component. They also generally specify the features every frame, rather than associating them with linguistic units because the features can

thus be automatically recognized from the acoustic signal prior to hypothesizing linguistic unit boundaries. This is in contrast to "landmark" approaches, which first hypothesize linguistically important events, and then produce either acoustic features for each event or distinctive features defined at landmarks (Section IV C).

The key advantage of using features in phonology transfers directly to statistical modeling. Features are a *factored* representation and, through this factorization, feature values are shared by several phonemes. As we already mentioned in the Introduction, even in standard HMM-based recognition systems (e.g., Young *et al.*, 2002), this factored representation is extremely useful.

Chomsky and Halle's features are an abstract representation of speech production. After all, they were used in a *phonological* theory, in which only symbolic processes (e.g., assimilation) are of interest. In speech modeling, we wish to represent more acoustic detail than this. Fortunately, the feature set can be used to describe some acoustic (non-phonological) processes. For example, we could describe a nasalized version of an English vowel by simply changing its nasal feature value from $-$ to $+$. A simple extension of this feature system (see Section V A 1) changes the interpretation of the feature values to be probabilistic, with values ranging from 0 to 1, thus allowing *degrees* of nasalization, for example.

1. Speech as articulatory gestures

A separate and also influential perspective on speech organization was developed at Haskins Laboratories during the 1980s (e.g., Browman and Goldstein, 1992). A central tenet of the gestural approach is that speech percepts fundamentally correspond to the articulatory *gestures* that produced the acoustic signal. Gestures typically involve several articulators working together in (loose) synchrony. In the gestural view of speech production, a "gestural score" is first produced, from which a task dynamic model (Saltzman and Munhall, 1989) generates articulatory trajectories. The score is written using a finite number of types of gesture, such as "bilabial closure" and "velic opening". An example gestural score, using Browman and Goldstein's *vocal tract variables*, is shown in Figure 1. These gestures correspond directly to physical actions of the articulators.

The gestural approach provides an account of variation in spontaneous or casual speech. Instead of using complex phonological rules to account for phenomena such as lenition, reduction and insertion, it uses simple and predictable changes in the relative timing of vocal tract variables (Browman and Goldstein, 1991). A vivid example of the representational power of the gestural approach is provided in Rubin and Vatikiotis-Bateson (1998) for the utterances "banana", "bandana", "badnana" and "baddata" where it is shown that the differences between these utterances all come down to differences in the timing of velar movement.

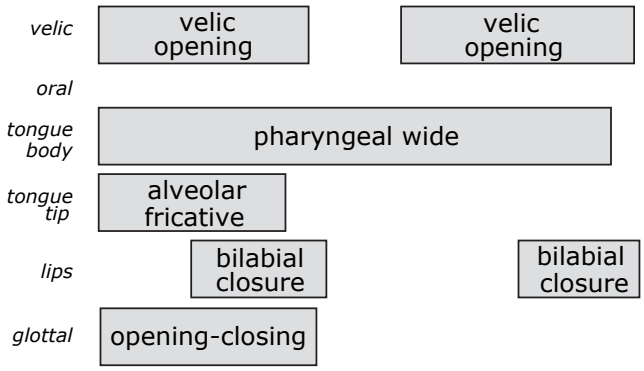

FIG. 1. Gestural score for the word "spam", adapted from Browman and Goldstein (1991).

C. Continuous-valued representations of speech production

In a few systems (e.g., Frankel, 2003; Wrench, 2001; Zlokarnik, 1995), a number of continuous-valued streams of data together give a smoothly-varying description of speech production. These may consist of measured human articulation, parameters automatically recovered from acoustic input (Section IV A), or production-*inspired* parameters (Nix and Hogden, 1998; Richards and Bridle, 1999).

The relationship between articulation and acoustics is highly non-linear. Measured articulatory parameters are smooth, slowly varying and continuous (yet noisy), due to the highly constrained dynamics of speech production. By contrast, acoustic parameters display regions of smooth spectral change interspersed by sudden discontinuities such as found in plosives. Analysis of speaker-dependent Electromagnetic Articulograph (EMA) data from the MOCHA corpus (Wrench and Hardcastle, 2000) reported by Frankel (2003) shows that a linear predictor can explain much of the variation found in articulatory parameters, both within and between phone segments. On acoustic data, whilst a linear predictor is suitable within phones, a non-linear predictor is required to model dependencies between phone segments.

III. MEASURING SPEECH PRODUCTION FROM THE ARTICULATORS

In this Section, we look at ways of measuring speech production, then in Section IV we will cover methods that only require the acoustics (at least, at recognition time). We must make a distinction between "data" and "knowledge". Whilst data (e.g., articulation measured by Electromagnetic Articulography or Magnetic Resonance Imaging) may be regarded as ground truth, it is not sufficient to build a *model*. Powerful machine-learning

techniques are available for learning the parameters of a model and, increasingly, for selecting amongst a set of candidate models. These techniques use *data* but require the *model* (or model type) to be specified. Machine-learning cannot hypothesize new types of model for us; for that, we must apply *knowledge* of the problem. In this article, we concentrate on systems that follow this methodology: they use knowledge of speech production to hypothesize a new type of model, then use machine-learning and data to learn its parameters.

A. Techniques for measuring articulation

Measuring articulation is frequently an invasive process. However, subjects generally manage to produce intelligible and reasonably natural speech despite the array of instruments to which they are attached.

An X-ray microbeam system involves attaching 2-3 mm gold pellets to the articulators which are tracked by a narrow, high-energy X-ray beam. The system described by Westbury (1994) achieves sampling rates of between 40 Hz and 160 Hz depending on the articulator being tracked. The X-ray machinery produces appreciable levels of background noise, resulting in a noisy speech signal and also interfering with speech production – the Lombard effect (Junqua, 1993).

An Electromagnetic Articulograph (EMA) system uses small receiver coils, instead of gold pellets, which have thread-like wires attached. These interfere surprisingly little with speech production. Current is induced in the coils by alternating magnetic fields from fixed transmitter coils mounted on a helmet, and their position can be inferred. As with an X-ray microbeam system, only x and y coordinates of each sensor in the *mid-sagittal plane* are measured, although a recently-developed three dimensional version overcomes this limitation and additionally removes the need to attach the transmitter coils to the subject (Hoole *et al.*, 2003). EMA systems can be located in recording studios, produce no operating noise, and therefore offer very high quality audio. In practice, the duration of recording sessions is limited because coils become detached or the subject tires.

A laryngograph, or electroglottograph (EGG), measures variation in conductance between transmitting and receiving electrodes positioned either side of the larynx, which is related to the change in glottal contact. An electropalatograph (EPG) measures tongue/palate contact over the whole palate using a custom-made artificial palate which has electrodes embedded on the lower surface in a grid pattern. Because it is a few millimeters thick, this interferes substantially with speech production. However, articulatory compensation (Perkell, 1997) means that relatively natural speech can be produced if the speaker wears the palate for some time before the recording session starts. An overview of other articulation-measuring devices can be found in Stone (1997), including computed tomography, magnetic resonance imaging, ultrasound, electromyography, strain gauges, video tracking systems, various aerodynamic measurement devices and so on.

B. Available corpora

Corpora large enough for training and evaluating ASR systems are scarce due to the expense and labor involved in data collection. We are aware of just two such data sets. The Wisconsin X-ray microbeam database (Westbury, 1994) consists of parallel articulatory and acoustic features for 60+ subjects, each of whom provide about 20 minutes of speech, including reading prose passages, counting and digit sequences, oral motor tasks, citation words, near-words, sounds and sound sequences, and read sentences. The MOCHA corpus (Wrench, 2001; Wrench and Hardcastle, 2000) was recorded at Queen Margaret University College, Edinburgh, and consists of parallel acoustic-articulatory recordings for a number of speakers, each of whom read up to 450 sentences from TIMIT (Garofolo *et al.*, 1993), plus 10 further sentences to cover the received pronunciation (RP) accent of British English. The measurements comprise EMA, EPG and EGG. Data sets from the other measurement devices listed above do exist, but not usually in useful quantities. New data are gradually becoming available, including EMA and magnetic resonance imaging (MRI) video data from the University of Southern California.[1]

IV. INFERRING SPEECH PRODUCTION FROM THE ACOUSTIC SIGNAL

In the absence of measurements of speech production, some method is required for recovering this information from the acoustic signal. Here, we discuss the tasks of articulatory inversion (Section IV A), articulatory feature recognition (Section IV B), and landmark detection (Section IV C). Articulatory inversion is concerned with faithful recovery of articulation or realistic, articulator-like parameters. Articulatory feature recognition is the inference of discrete pseudo-articulatory states. Landmark detection aims to enhance feature detection by locating points in the signal at which reliable acoustic cues may be found.

A. Articulatory inversion

An inversion mapping seeks to invert the process of speech production: given an acoustic signal, it estimates the sequence of underlying articulatory configurations. There is evidence that multiple articulatory configurations can result in the same or very similar acoustic signals: a many-to-one mapping. This makes the inversion mapping one-to-many, which is an *ill-posed* problem. For example, using the Wisconsin University X-ray microbeam database, Roweis (1999) showed that the articulatory data points associated with the nearest thousand acoustic neighbors of a reference acoustic vector could be

spread widely in the articulatory domain, sometimes in multimodal distributions.

One method of inversion uses an analysis of the acoustic signals based on some mathematical model of speech production and the physical properties of the articulatory system (Krstulović, 1999; Wakita, 1979). Another technique uses articulatory speech synthesis models with an analysis-by-synthesis algorithm: model parameters are adjusted so the synthesizer output matches the acoustic target (e.g., Shirai and Kobayashi, 1986). Synthesis models can be used to generate articulatory-acoustic databases, which can then be used for performing the inversion mapping as part of a code-book inversion method (e.g., Atal *et al.*, 1978) or as training data for another data-driven machine-learning model (e.g., Rahim *et al.*, 1993). A fundamental problem facing the use of analytical methods and of articulatory synthesis models is the difficulty in evaluating the result with respect to real human articulation. From this point of view, measurements of human articulation can provide a huge advantage. (Section VI E considers the problem of evaluation more generally).

Together with new data, the popularity of machine-learning methods has led to a recent increase in data driven methods, including extended Kalman filtering (Dusan and Deng, 2000), self-organizing HMMs (Roweis, 1999), codebooks (Hogden *et al.*, 1996), artificial neural networks (ANNs) such as the Multilayer Perceptron (Papcun *et al.*, 1992; Richmond *et al.*, 2003) and the Mixture Density Network (Richmond *et al.*, 2003), Gaussian Mixture Models (Toda *et al.*, 2004), and an HMM-based speech production model (Hiroya and Honda, 2004).

Finally, there are approaches that do not rely on *articulatory* data: so-called latent variable models, such as the Maximum Likelihood Continuity Map (MALCOM) (Hogden *et al.*, 1998), which estimates the most likely sequence of hidden variables in accordance with articulatory-like constraints, given a sequence of acoustic frames. The constraint is simply that the estimated motion of pseudo-articulators cannot contain frequency components above a certain cutoff frequency, e.g. 15 Hz. This is a direct use of the knowledge that articulator motion is smooth (more specifically, band-limited) to aid the inversion process.

Although Hogden *et al.* do not use articulatory data for training, they report that the trajectories of the hidden variables correlate highly with measured articulatory trajectories.

B. Articulatory feature recognition

Articulatory feature recognition can be incorporated directly into existing phone or word-based systems, or can provide a subtask on the way to building a full AF-based ASR system. Typically, separate models are trained to distinguish between the possible values of each feature. Kirchhoff (1999) proposes this approach because the complexity of each individual classifier will be less than that of a monolithic classifier, leading to improved robustness. Efficient use is made of training data, improving the modeling of infrequently occurring feature combinations. ANNs, HMMs, and dynamic Bayesian networks (DBNs) have all successfully been applied to the task of AF recognition.

Attempting the task of AF recognition, without actually performing word recognition, presents two inherent difficulties, both of which stem from deriving AF labels from phone labels: obtaining labels for the data and evaluating the system. If the AFs directly correspond to articulator positions, then they may be obtained by quantizing articulatory measurement data. If AF labels are produced from phonetic transcriptions, there is the possibility of merely having a *distributed* representation of phonemes without the advantages of a truly *factored* representation. Embedded training (e.g., Wester *et al.*, 2004) can be used to address limitations in phone-derived AF labels, by allowing boundaries to be moved and labels potentially changed.

The following work on AF recognition all aims towards full ASR; AF recognition is merely a staging post along the way. There are two distinct categories of work here. The first uses AFs of the kind discussed above; we describe three different machine-learning approaches to recognizing the values of AFs from speech. The second approach is that of landmarks. Evaluation of such systems can present some problems, which are discussed in Section VI E.

1. Articulatory feature recognition using neural networks

King and Taylor (2000) report articulatory feature recognition experiments on TIMIT using ANNs, comparing three feature systems: binary features based on the *Sound Pattern of English* (SPE) (Chomsky and Halle, 1968), multivalued features using traditional phonetic categories such as manner, place, etc., and *Government phonology* (GP) (Harris, 1994). The percentage of frames with all features simultaneously correct together was similar across feature systems: 52%, 53% and 59% for SPE, multivalued and GP respectively (59%, 60% and 61% when each frame was mapped to the nearest phoneme). Dalsgaard *et al.* (1991) aligned acoustic-phonetic features with speech using a neural network; the features were similar to the SPE set, but underwent principal components analysis to reduce correlation. It is not clear whether they retain linguistic meaning after this procedure. Kirchhoff (Kirchhoff, 1999; Kirchhoff *et al.*, 2002) used articulatory features to enhance a phone-based system. Wester *et al.* (2001) and Chang *et al.* (2005) used separate place classifiers for each value that manner can take. Omar and Hasegawa-Johnson (2002) used a maximum mutual information approach to determine subsets of acoustic features for use in AF recognition.

2. Articulatory feature recognition using hidden Markov models

A number of systems use HMMs for AF recognition, including Metze and Waibel (2002), who used the set of linguistically motivated questions devised for clustering context-dependent HMM phone models to provide an initial AF set. A set of feature detectors was then used to supplement an HMM system via likelihood combination at the phone or state level. Word error rate (WER) was reduced from 13.4% to 11.6% on a 40k word vocabulary Broadcast News task and from 23.5% to 21.9% on spontaneous speech from the Verbmobil task. An HMM approach was also taken by Eide (2001), and used to generate observations for further processing in a phone-based HMM system.

3. Articulatory feature recognition using dynamic Bayesian networks

In contrast to the use of ANNs and HMMs, the use of DBNs is motivated specifically by their particular capabilities for this task: the ability to transparently and explicitly model inter-dependencies between features and the possibility of building a single model that includes both AF recognition and word recognition.

Frankel *et al.* (2004) proposed DBNs as a model for AF recognition. As with the manner-dependent place ANNs discussed above, evaluation on the TIMIT corpus (Garofolo *et al.*, 1993) showed that modeling inter-feature dependencies led to improved accuracy. The model is shown in figure 2. Using phone-derived feature labels as the gold standard, the overall framewise percent features correct was increased from 80.8% to 81.5% by modeling dependencies, and frames with all features simultaneously correct together increased dramatically from 47.2% to 57.8% (this result can be compared to 53% for King and Taylor's multivalued feature system described in Section IV B 1, where the feature system was very similar).

To mitigate the problems of learning from phone-derived feature labels, an embedded training scheme (mentioned in Section IV B) was developed by Wester *et al.* (2004) in which a set of asynchronous feature changes was learned from the data. Evaluation on a subset of the OGI Numbers corpus (Cole *et al.*, 1995) showed that the new model led to a slight increase in accuracy over a similar model trained on *phone-derived* labels (these accuracy figures do not tell the whole story – see Section VI E 2 a). However, there was a 3-fold increase in the number of feature combinations found in the recognition output, suggesting that the model was finding some asynchrony in feature changes. Frankel and King (2005) describe a hybrid ANN/DBN approach, in which the Gaussian mixture model (GMM) observation process used by the original DBNs is replaced with ANN output posteriors. This gives a system in the spirit of hybrid ANN/HMM speech recognition (Bourlard and Morgan, 1993), combining the benefit of the ANN's discriminative

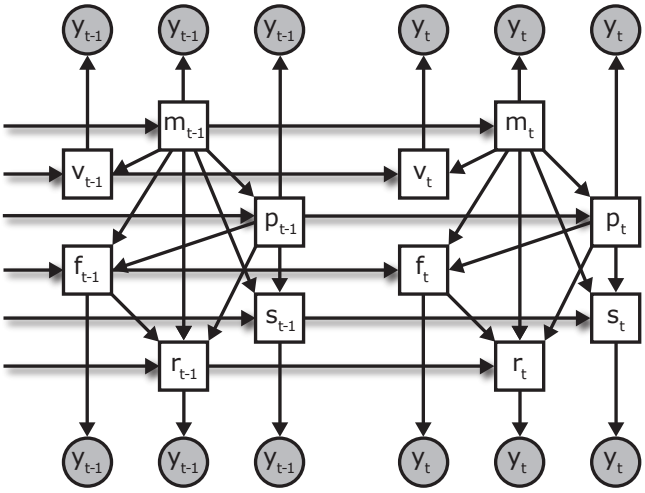

FIG. 2. A DBN for articulatory feature recognition, from Frankel *et al.* (2004), shown in graphical model notation where square/round and shaded/unshaded nodes denote discrete/continuous and observed/hidden variables respectively. (Arcs between time slices are drawn here, and in all other DBN figures, with "drop shadows" for clarity, although they are otherwise no different to other arcs.) The discrete variables $m_t, p_t, v_t, f_t, s_t, r_t$ are the articulatory features (manner, place, voicing, frontback, static, rounding) at time t. The model consists of 6 parallel HMMs (e.g., consider only the variables m_t, m_{t-1} and their corresponding observations) plus a set of inter-feature dependencies (e.g., the arc from m_t to p_t indicates that the distribution of p_t depends on the value of m_t). These inter-feature dependency arcs allow the model to learn which feature values tend to co-occur. The continuous observation y_t is repeated 6 times (a product-of-Gaussians observation density).

training with the inter-feature dependency modeling offered by the DBN. The feature recognition accuracy on OGI Numbers was increased to 87.8%.

C. Landmark-based feature detection

Feature-based representations have been used for a long time in the landmark-based recognition work of Stevens (2000; 2002) which models speech perception in humans; this work has inspired landmark-based ASR (e.g., Hasegawa-Johnson *et al.*, 2005; Juneja, 2004; Zue *et al.*, 1989). Stevens (2002) describes the recognition process as beginning with hypothesizing locations of *landmarks*, points in the speech signal corresponding to important events such as consonantal closures and releases, vowel centers and extrema of glides (e.g., Figure 3). The type of landmark determines the values of *articulator-free* features such as [sonorant] and [continuant]. Various cues (e.g., formant frequencies, spectral amplitudes, duration of frication), are then extracted around the landmarks and used to determine the values of *articulator-bound* distinctive features (e.g., place,

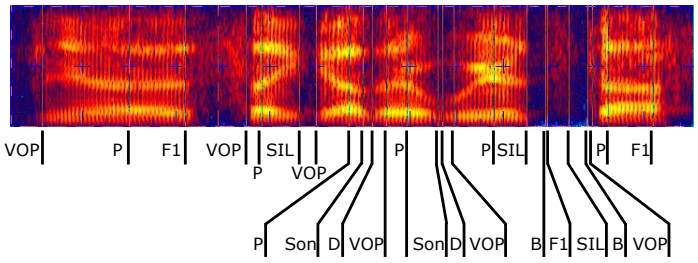

FIG. 3. Landmarks for the utterance "Yeah it's like other weird stuff." Text labels correspond to the landmark to their right. F1: fricative onset; Son: sonorant consonant onset; P: vowel nucleus; D: syllabic dip; SIL: silence onset; B: stop burst onset; VOP: vowel onset point. Based on Figure 4.2 in Hasegawa-Johnson *et al.* (2005).

vowel height, nasality). The set of hypothesized features is matched against feature-based word representations in the lexicon.

There is no system for automatic speech recognition of which we are aware that implements this theory fully, but modules implementing some aspects of the theory have been developed. Howitt (1999) reports on a method for vowel landmark detection using a simple multilayer perceptron; Choi (1999) presents knowledge-based cues and detectors for consonant voicing at manually-labeled landmarks. Two complete ASR systems using landmarks – the MIT SUMMIT system and a system from the 2004 Johns Hopkins Summer Workshop – are described in Section V B. Juneja and Espy-Wilson (Juneja, 2004; Juneja and Espy-Wilson, 2003b) report related work, which combines support vector machine outputs with dynamic programming to locate landmarks and label manner features.

D. Comparison of AF and landmark approaches

In most AF-based systems, AF values are defined every time frame. This is compatible with any of the frame-based modeling techniques described in the next section, including generative models such as HMMs. These frame-based models compute the likelihood of an utterance given some hypothesized labels (e.g., a word sequence) by multiplying together frame-level likelihoods. It is also straightforward to compute a frame-level accuracy for these types of systems, so long as reference labels are available (Section VI E discusses this issue). In contrast, landmarks are *events*. Evaluation of landmark accuracy requires a measure such as the F score which combines recall and precision, *plus* some measure of the temporal accuracy of the landmarks. For this reason, reports of landmark accuracy are less common, less consistent, and harder to interpret. For subsequent word recognition, landmarks are used to guide acoustic feature extraction, meaning that these acoustic features are not available at every time frame to any subsequent model.

V. ACOUSTIC MODELING USING PRODUCTION KNOWLEDGE

We now consider how speech production knowledge has been used to improve acoustic modeling for speech recognition. Some of the work builds on AF recognition described above. A simple way to use articulatory features is as a replacement for conventional acoustic observations. Alternatively, "landmarks" may be located in the signal, and acoustic parameters extracted at those locations. AFs can be used to perform phone recognition, where they have been shown to improve noise robustness, although this approach suffers from the phone "bottleneck" that AF approaches usually try to avoid. A rather different way to harness the power of articulatory information is to use it for model selection by, for example, defining the topology of an otherwise conventional HMM. Other models maintain an explicit internal representation of speech production, whether that be discrete or continuous. AFs have also been used in pronunciation modeling, and for recognition-by-synthesis using an articulatory speech synthesizer.

A. Articulatory features or parameters as observations

Articulatory parameters (continuous-valued or quantized) can be used directly as (part of) the observation vector of a statistical model. This requires access to measured or automatically-recovered articulation.

1. Hidden Markov Models

Zlokarnik (1995) used measured or automatically-recovered articulatory parameters, appended to acoustic features, in an HMM recognizer. On VCV sequences, adding measured articulation to Mel-frequency cepstral co-efficients (MFCCs) reduced WER by more than 60% relative. Articulatory parameters recovered using a multilater perceptron (MLP) gave relative WER reductions of around 20%. The additional information carried by the recovered articulation may be attributable either to the supervised nature of MLP training, or the use of 51 frames (approximately half a second) of acoustic context on the MLP inputs.

Similar experiments were conducted on a larger scale by Wrench. For a single speaker, augmenting acoustic features with measured articulatory parameters gave a 17% relative phone error rate reduction using triphone HMMs. The articulatory feature set was generated by stacking EMA, EGG and EPG signals with their corresponding δ (velocity) and $\delta\delta$ (acceleration) coefficients and performing linear discriminant analysis (LDA) dimensionality reduction. However, when real articulation was replaced with articulatory parameters automatically recovered from the acoustics using an MLP, there was no improvement over the baseline acoustic-only result (Wrench, 2001; Wrench and Hardcastle, 2000; Wrench

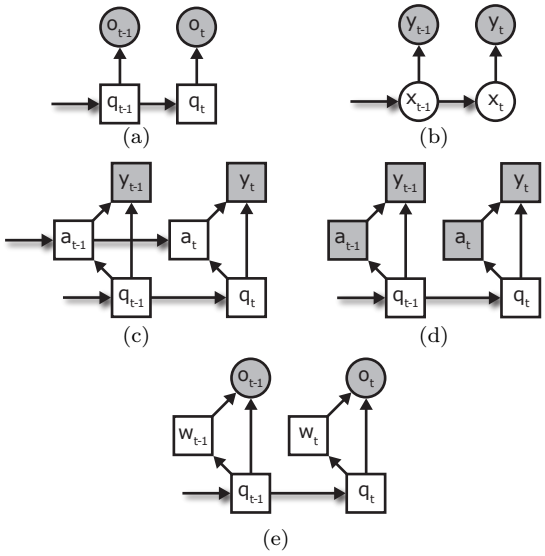

FIG. 4. 4(a): HMM in graphical model notation. q_t is the hidden discrete state at time t and o_t is the continuous-valued observation. This graph illustrates the conditional independence assumptions in this model, e.g., o_t is conditionally independent of everything except q_t, given q_t. 4(b): Linear dynamic model (LDM). x_t is the continuous hidden state and y_t is the continuous observation at time t. The model is similar to the HMM except the state is now continuous: its value is a vector of real numbers. Whereas the HMM stochastically chooses a state sequence, the LDM makes stochastic trajectories in its state-space. 4(c): DBN from Stephenson *et al.* (2000). q_t is the hidden state at time t, y_t is a discrete acoustic observation and a_t is a discretized articulator position (which may be observed during training). This model is somewhat similar to the conventional HMM with a mixture-of-Gaussians output density in Figure 4(e), except that the observation is discrete (for practical implementation reasons). The dependency of a_t on a_{t-1} models the dynamics of articulator movement. 4(d): Hybrid HMM/BN from Markov *et al.* (2003).. q_t is the hidden state at time t, y_t is a discrete acoustic observation and a_t is an observed, discretized articulator position. The model is very similar to that of Stephenson except the articulator position is always observed (so the dependency of a_t on a_{t-1} is not needed). 4(e): A conventional mixture-of-Gaussians continuous-density HMM. w_t is the hidden mixture component at time t (the probability mass function of w_t is constant: it is the set of mixture weights). The production-based models in Figures 4(c) and 4(d) use a similar structure to achieve a mixture distribution over the observation.

and Richmond, 2000).

Eide (2001) describes augmenting the MFCC observation vector for a standard HMM system with information about articulatory features. Mutual information between the true and estimated presence of features was used to reduce the original 14 features down to 4. In an evaluation on city and street names spoken in a car with the engine running at 0, 30 and 60 mph, the augmented observations gave 34%/22% relative word/string error rate reductions.

Fukuda *et al.* (2003) used an MLP to map from acoustic parameters to a 33-dimensional output vector representing 11 distinctive phonetic features (DPFs) at the current time frame along with inferred values at preceding and following contexts. The modeling of the MLP-derived DPFs was refined through the application of logarithmic feature transforms (Fukuda and Nitta, 2003a) and dimensionality reduction (Fukuda and Nitta, 2003b). Augmenting MFCC feature vectors with the MLP-derived DPFs gave accuracy increases over the baseline, particularly in the presence of noise.

King *et al.* (1998) and King and Taylor (2000) also report recognition experiments based on the combination of the output of a number of independent ANN classifiers. The work was primarily aimed at comparing phonological feature sets on which to base the classifiers, though the feature predictions were also combined to give TIMIT phone recognition results. Unlike Kirchhoff, who used an ANN to combine the independent feature classifiers, the predicted feature values were used as observations in an HMM system. The resulting recognition accuracy of 63.5% was higher than the result of 63.3% found using standard acoustic HMMs, though not statistically significant. The need for an asynchronous articulatory model was demonstrated using classifications of a set of binary features derived from Chomsky and Halle (1968). In cases where features changed value at phone boundaries, allowing transitions within two frames of the reference time to be counted as correct, the percentage of frames where all features were correct (i.e., where the phone was correctly identified) rose from 52% to 63%. Furthermore, the accuracy with which features were mapped onto the nearest phone rose from 59% to 70%. This demonstrates the limiting nature of forcing hard decisions at phone boundaries onto asynchronous data. In both King *et al.*'s and Kirchhoff's systems, the individual feature classifiers were independent.

2. *Hybrid HMM/ANN systems*

Kirchhoff (Kirchhoff, 1999; Kirchhoff *et al.*, 2002) showed that an AF-based system can increase robustness to noise. The OGI Numbers corpus (Cole *et al.*, 1995) was used to develop this approach, using the feature representation given in Section II B. Feature labels were generated from time-aligned phone labels using rules. As in Section IV B, a separate MLP for each feature was trained to estimate posterior probabilities, given the acoustic input. A further MLP was trained to map from the outputs of the 5 feature networks to phone class posteriors which were then used in a standard hybrid HMM/ANN recognition system.

On clean speech, the word error rates for the acoustic and articulatory models were comparable, 8.4% and 8.9% respectively, though in the presence of a high degree of additive noise, the articulatory model produced significantly better results. At 0 dB (signal and noise have equal intensity), the word error rate for the acoustic model was 50.2% but was 43.6% for the articulatory system. When the outputs of the acoustic and articulatory recognizers were combined, the error rates were lower than for either of the two individually, under a variety of noise levels and also on reverberant speech.

The framewise errors for the different articulatory feature groups showed that classification performance on the voicing, rounding and front-back features do not deteriorate as quickly as for manner and place in the presence of noise. This result suggests that, by incorporating confidence scores when combining the outputs of individual classifiers, the system could be tailored to particular operating conditions, and supports the authors' suggestion that combining individual classifiers might lead to improved robustness over a monolithic classifier (i.e., one that recognizes all features simultaneously). Similar experiments were performed on a larger spontaneous dialog corpus (Verbmobil). Improvements were also shown when acoustic and articulatory features were combined, giving relative WER reductions of up to 5.6%.

3. Dynamic Bayesian Networks

Stephenson *et al.* (2000, 2004) created a DBN which enhances the output mixture distribution of an HMM by including dependency on an articulator position. Figure 4(c) shows two time-slices of the model. The articulator position is conditioned on its previous value and on the current sub-word state, providing an element of contextual modeling. Note that the decoding version of the model is shown in which the articulator position is hidden. During training, the articulator position may be observed. The Wisconsin X-ray microbeam database (Westbury, 1994) was used to provide parallel acoustic-articulatory data for an isolated word recognition task. The acoustic features were 12 MFCCs and energy along with their δ coefficients, and the articulatory features consisted of x and y coordinates for 8 articulator positions (upper lip, lower lip, four tongue positions, lower front tooth, lower back tooth). Both acoustic and articulatory observations were discretized by generating codebooks using K-means clustering. The acoustic-only word error rate of 8.6% was reduced to 7.6% when the articulatory data was used during recognition. With the articulation hidden, the system gave a recognition word error rate of 7.8%, which is a 9% relative error decrease over the acoustic baseline.

4. Hybrid Hidden Markov Model/Bayesian Network

Similar work is described in Markov *et al.* (2003), using a hybrid HMM plus Bayesian network (BN): an HMM in which the BN provides the observation process. The hybrid HMM/BN shown in Figure 4(d) was used to implement a similar system to that in Stephenson *et al.* (2000), but without the dependency between successive articulator positions. By conditioning the GMM observation distributions on both the sub-word state and the (discrete) articulator value, the model is an HMM with a mixture output distribution where the mixture component (i.e., articulator position) is observed. Figure 4(e) shows a standard HMM with continuous observations and a mixture-of-Gaussians output density for comparison. As above, real articulatory data collected on an EMA machine was used for training the models, with the data first discretized. Unlike Stephenson *et al.* (2000), continuous-valued acoustic observations were used. Speaker-dependent experiments showed that the structure in the BN observation process makes it possible to support more mixture components than with standard GMMs (Markov *et al.*, 2003). Using 300 training sentences of parallel acoustic and articulatory data from 3 speakers, HMMs trained and tested on both acoustic and articulatory data significantly outperformed HMMs trained and tested on only acoustic data. HMM/BN models trained on both acoustic and articulatory data, even though performing recognition using only the acoustic parameters, gave similar performance to the HMMs trained and tested on both. These findings support those of Stephenson *et al.* (2000). Both these systems require articulatory measurement data for training.

5. Linear Dynamic Models

In the preceding work using HMMs and HMM/BNs, no attempt (other than the use of delta features) was made to model the continuous nature of articulator positions through time. Only Stephenson's model includes a dependency between the current articulatory state variable and its value at the preceding time, but this variable is discrete. All of these models use only discrete hidden state variable(s). In contrast, linear dynamic models (LDMs) use a *continuous* state variable.

Frankel *et al.* (Frankel, 2003; Frankel and King, 2001a,b; Frankel *et al.*, 2000) report the results of phone classification and recognition on the MOCHA corpus using LDMs. These are generative state-space models in which a continuous-valued hidden state variable gives rise to a time-varying multivariate Gaussian output distribution. Figure 4(b) shows two time-slices of a LDM in graphical model notation. Frankel (2003) describes a phone classification task comparing various types of observation vectors derived from the MOCHA corpus. These include MFCCs, measured articulation (EMA), EGG and EPG data. Acoustic-only observations gave higher accuracy than EMA alone, but when EGG (i.e. voicing information) and EPG data was added to EMA, the accuracy approaches that of the acoustic-only system. The acoustic-only phone error rate was reduced by 16.2% relative by adding EMA. Replacing measured EMA parameters with values recovered from the acoustics by an MLP (Richmond *et al.*, 2003) actually led to a slight reduction in accuracy, compared to the acoustic-only system. This may be due to the type of feed-forward MLP used in the inversion mapping, which estimates the conditional average articulatory parameters, given the acoustic inputs. Papcun *et al.* (1992) and Rose *et al.* (1996) observed that non-critical articulators tend to have higher associated variance than critical articulators. With no provision to model this variation, the MLP will intro-

duce consistency where there should be none which may lead to an overemphasis on data streams corresponding to non-critical articulators. An alternative type of network might be better: in Richmond (2002) a mixture density network was applied to the inversion task. Such networks can model one-to-many relationships and account for variance, because their outputs are mixtures of probability density functions.

B. Landmark-based systems

The idea of locating landmarks in the speech signal, and using those locations to extract acoustic information, was introduced in Section IV C. The use of landmarks does not, in itself, imply the use of production knowledge or articulatory features, and has been used as part of both phone-based and articulatory feature-based recognition systems.

The MIT SUMMIT speech recognition system (Glass, 2003; Zue *et al.*, 1989) formalizes some of the ideas of Stevens' landmark-based approach (Stevens, 2002) in a probabilistic setting. SUMMIT locates potential phone-boundary landmarks and uses a phone-based dictionary to represent words. SUMMIT has used various landmark detection algorithms (Chang and Glass, 1997; Glass, 1988) and acoustic cues (Halberstadt and Glass, 1998; Muzumdar, 1996; Zue *et al.*, 1989). SUMMIT operates in either (or both) of two modes: a boundary-based mode, in which the acoustic cues around phonetic-boundary landmarks are explicitly modeled, and a segment-based mode, in which the regions between landmarks are modeled. Recent work by Tang *et al.* (2003) uses SUMMIT in a combined phone-feature approach to word recognition.

The 2004 Johns Hopkins Summer Workshop project on landmark-based speech recognition used an entirely feature-based representation of words rather than a phonetic one (Hasegawa-Johnson *et al.*, 2005). It also differed from SUMMIT in that it used support vector machines (SVMs) to detect both landmarks and the presence or absence of distinctive features. The outputs of these SVMs were combined into word scores and used to rescore word lattices produced by a baseline HMM-based recognizer. This project experimented with three ways of combining the SVM outputs into word scores. The first system used the approach of Juneja and Espy-Wilson (2003a), in which SVM discriminant scores are converted to likelihoods and modified Viterbi scoring is done using a phonetic baseform dictionary, mapped to distinctive features. The second system used an articulatory feature-based pronunciation model inspired by that of Livescu and Glass (Section V E) and the third used a maximum entropy model to classify words in a confusion network.

C. Articulatory features for HMM model selection

Articulatory features may also be used for the purposes of model selection, providing a prior on model topology by specifying the function of sub-word states. This is distinct from AFs providing the internal representation because, in the model selection approach, once the model is selected (e.g., the topology of an HMM is specified), the articulatory information is no longer required.

1. Feature bundles

Deng and colleagues (e.g., Deng and Sun, 1994a,b; Sun *et al.*, 2000) have developed HMM systems where each state represents an articulatory configuration. Following Chomsky and Halle's theory of distinctive features and Browman and Goldstein's system of phonology (Browman and Goldstein, 1992), they developed a detailed system for deriving HMM state transition networks based on a set of 'atomic' units. These units represent all combinations of a set of overlapping articulatory features that are possible under a set of hand-written rules. Each phone is mapped to a static articulatory configuration (affricates and diphthongs each have a sequence of two configurations). Features can spread, to model long span dependencies. When articulatory feature bundles overlap asynchronously, new states are created for the intermediate portions which describe transitions or allophonic variation. On a TIMIT classification task, HMMs constructed from these units achieved an accuracy of 73% compared with context-independent HMMs of phones which gave an accuracy of 62%. The feature-based HMMs also required fewer mixture components. This suggests that a principled approach to state selection requires fewer parameters and therefore less training data, since each state is modeling a more consistent region of the acoustic space. This work was extended to include higher level linguistic information (Sun *et al.*, 2000), including utterance, word, morpheme and syllable boundaries, syllable onset, nucleus and coda, word stress and sentence accents. This time, results were reported on TIMIT phone recognition, rather than classification. A recognition accuracy of 73% was found using the feature-based HMM, which compares favorably to their baseline triphone HMM which gave an accuracy of about 71%, although this is not a state-of-the art accuracy.

2. Hidden articulator Markov model

Richardson *et al.* (2000a,b) drew on work by Erler and Freeman (1996) in devising the hidden articulator Markov model (HAMM), which is an HMM where each articulatory configuration is modeled by a separate state. The state transitions reflect human articulation: Static constraints disallow configurations which would not occur in American English, and dynamic constraints ensure that only physically possible movements are allowed. Asynchronous articulator movement is allowed: Each feature can change value independently of the others. On a 600 word PHONEBOOK isolated word, telephone speech, recognition task, the HAMM gave a significantly higher WER than a 4-state HMM (7.56% vs. 5.76%) but

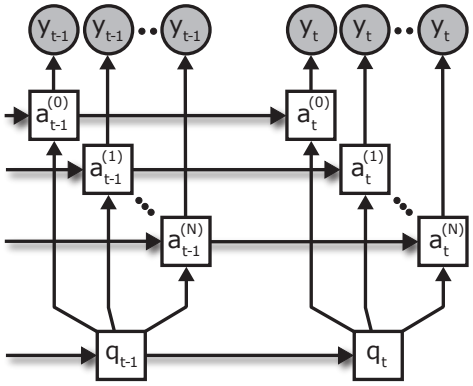

FIG. 5. A hidden feature model from Livescu *et al.* (2003) shown in graphical model notation. q_t is the (phonetic) hidden state, y_t is the acoustic observation and $a_t^{(0)} \dots a_t^{(N)}$ are the articulatory variables, at time t. The dependency of the observation on the phone is mediated by the articulatory variables. Adding this intermediary layer allows for feature-based pronunciation modeling via the phone-to-feature dependencies.

a combination of the two gave a WER 4.56%: a relative reduction of 21% over the HMM system.

D. Articulatory information as internal structure

Articulatory information can be used to provide some or all of the internal model structure. This can take the form of decomposing sub-word states into a set of discrete articulatory features (Bilmes *et al.*, 2001; Livescu *et al.*, 2003), or using a continuous-valued articulatory-like representation which then generates acoustic parameters via some transform (Iso, 1993; Richards and Bridle, 1999; Russell and Jackson, 2005); some of the linear dynamic models of Frankel and colleagues (Section V A 5) can be seen as having a hidden continuous articulatory-like state variable which generates acoustic observations.

1. Factoring the state into discrete articulatory features

Features provide a parsimonious framework within which to represent the variation present in natural speech. The approaches below use AF recognition (Section IV B 1) coupled with some method of mapping from features to sub-word or word units; thus, the AFs are explicitly present in the model's internal representation.

Eide *et al.* (1993) presents work in this vein, though phones are used to mediate between features and words, which compromises the benefits of a feature approach. Kirchhoff (1996) observes that articulatory asynchrony spans units longer than phones, and describes a system in which synchronization of feature streams is delayed to the syllable level. HMMs are used to infer values corresponding to each of 6 feature streams, and syllables are defined as parallel sequences of feature values. In evaluation on spontaneous German speech, a baseline triphone-based recognizer gave a phone accuracy of 54.81%. To allow comparison, recognized syllables were mapped to a phone sequence, and gave the substantially higher recognition accuracy of 68.3%, although it should be noted that this not a fair comparison because the syllable-based system benefits from the phonotactic constraints provided by the syllable models.

Bilmes *et al.* (2001) proposed a DBN-based approach to ASR with an AF internal representation. Livescu *et al.* (2003) continued this work and proposed a model that uses an articulatory feature factorization of the state space. A set of 8 features is defined, with the value of each conditioned on the current phone state and its own previous value (Figure 5). Dependencies may also be added between features in the same time-slice. To overcome the problem of specifying an observation model for every possible combination of features, a product-of-mixtures-of-Gaussians model is used. Evaluation of the model on the Aurora 2.0 noisy digit corpus showed small accuracy increases over a phone-based HMM baseline in clean test conditions and more substantial improvements in some noise conditions. However, improvements over the baseline were only found when a phone-to-observation edge was included, giving a system in which the feature- and phone-based model likelihoods are effectively combined at the frame level. Only limited forms of such models were considered (for computational reasons), in which the inter-frame feature dependencies shown in Figure 5 were omitted and features were conditionally independent of each other. Given the flexibility of the DBN framework, there is much scope for further development of this approach.

There have been other attempts to use a factored state representation. For example, Nock (2001) proposed "Loosely-coupled HMMs" which have two or more Markov chains, each with its own observation variable. In Nock's work, the observations for each chain were derived from a different frequency band of the spectrum. Although it is clear that the acoustic consequences of speech production do not factor neatly into frequency bands, Nock's approach is inspired by the asynchronous nature of speech production and the loosely-coupled HMM may be more effective with observation streams that relate more directly to speech production (e.g., articulatory features).

2. Continuous articulatory internal representation

A number of researchers have investigated the use of continuous state-space representations, where the acoustic observations are modeled as the realization of some (possibly unobserved) dynamical system. Some of these approaches, such as the linear Gaussian systems described by Digalakis (1992), Frankel (2003), and Rosti (2004) are intended to reflect only the *general properties* of speech production and provide a compact representa-

tion of acoustic parameters (Section V A 5). Other studies, as described below, make a more explicit attempt to incorporate a model of the relationship between articulatory and acoustic domains.

a. Segmental HMM Russell and Jackson (2005) describe a multi-level segmental hidden Markov model (MSHMM) in which formant frequencies are used to build an articulatory-like internal representation. Each state in the model generates a variable-duration noisy linear trajectory in articulatory space, which is projected into the acoustic space via a linear mapping. The articulatory-acoustic mappings are either per-phone basis or shared across phone classes. A number of tying strategies were compared, with greater numbers of mappings giving improved performance. Given the linear nature of the articulatory trajectories, and the linear mapping to the acoustic parameters, a theoretical upper bound on performance is given by a fixed linear-trajectory acoustic segmental HMM (FT-SHMM) (Holmes and Russell, 1999), which models the acoustic parameters directly. Experimental results show that this bound is met, and that where triphone models are employed, the MSHMM gives comparable performance to the FT-SHMM system with a 25% reduction in the number of free parameters. To overcome the limitations of using a linear mapping between articulatory and acoustic domains, Jackson *et al.* (2002) investigated the non-linear alternatives of MLP and radial basis function (RBF), finding superior performance with the RBF (for background reading on these and other machine-learning techniques, see Section VI G).

b. Long-span contextual modeling Attempting to capture long-span contextual effects along with the non-linear relationship between articulatory and acoustic parameters has prompted models such as that of Iso (1993) and the hidden dynamic model (HDM) of Richards and Bridle (1999). The HDM uses a segmental framework in which a static target or series of targets in a hidden state space is associated with each phone in the inventory. A Kalman smoother is run over the targets to produce a continuous trajectory through the state space. These trajectories are connected to the surface acoustics by a single MLP. For an N-best rescoring task on the Switchboard corpus and a baseline WER of 48.2% from a standard HMM system, 5-best rescoring with the reference transcription included[2] using the HDM gave a reduced error rate of 34.7% (Picone *et al.*, 1999). An identical rescoring experiment using an HMM trained on the data used to build the HDM gave a word error rate of 44.8%. This suggests that the HDM was able to capture information that the HMM could not.

Deng and Ma (2000) describe a similar model in which the state is intended to model the pole locations of the vocal tract frequency response via vocal-tract-resonance (VTR) targets for each phone. Multiple switching MLPs are used to map from state to observations, though instead of the deterministic output distribution found in the HDM, filtering is implemented with an extended Kalman filter (EKF). To avoid the difficulties of training a non-linear mapping and the inherent problems of the EKF, Ma and Deng (2004a,b) describe a system in which a mixture of linear models is used to approximate the non-linearity, and demonstrate slight error reductions over an HMM baseline on a Switchboard rescoring task.

Zhou *et al.* (2003) describe a hidden-trajectory HMM (HTHMM) which also combines VTR dynamics with a mixture of linear projections to approximate a non-linear state-to-observation mapping. However, the model is frame-based rather than segmental, and the state trajectories are deterministic, conditioned on the sequence of sub-word units, which in fact consist of HMM states. The model can be interpreted as an HMM in which the output distributions are adapted to account for long-span contextual information by conditioning on a continuous hidden trajectory. The deterministic continuous state obviates the need for filtering to infer state trajectories that, in combination with frame-based computation, simplifies decoder implementation – described in Seide *et al.* (2003). Initial evaluation on TIDIGITS (Leonard, 1984) with a context-independent HTHMM system produced 0.37% WER and matched the 0.40% WER of a context-dependent triphone HMM system.

E. Articulatory feature modeling of pronunciation variation

The usual choice of sub-word unit is the phoneme and the usual representation of a word is as a string of phonemes. AFs are an alternative to phonemes and their use is motivated by difficulties in describing pronunciation variation using a string of phonemes. Spoken pronunciations often differ radically from dictionary baseforms, especially in conversational speech (Weintraub *et al.*, 1996). This contributes to the poor performance of ASR (Fosler-Lussier, 1999; McAllaster *et al.*, 1998; Saraclar *et al.*, 2000). Phoneme-based pronunciation models usually account for variability by expanding the dictionary with additional pronunciation variants (Hazen *et al.*, 2005; Riley and Ljolje, 1996; Shu and Hetherington, 2002; Wester, 2003). However, phoneme-based pronunciation models have numerous drawbacks. Saraclar *et al.* (2000) show that a phonetic realization is often somewhere between the intended phoneme and some other phoneme, rather than a phonemic substitution, insertion, or deletion. Phonemic changes can lead to increased confusability; e.g., "support" will be confusable with "sport" if it is allowed to undergo complete deletion of the schwa. In reality though, the [p] in "support" will be aspirated even if the schwa is deleted; the one in "sport" will not. Bates (2003) addresses these drawbacks by building a model of phonetic substitutions in which the probabilities of possible realizations of a phoneme are computed using a product model in which each product

term involves a different AF or AF group.

Livescu and Glass (Livescu, 2005; Livescu and Glass, 2004a,b) generate pronunciation variants from baseforms through feature substitution and feature asynchrony using features based on Browman and Goldstein's vocal tract variables (Browman and Goldstein, 1992). Effects that can be modeled include: *asynchrony only*: Nasal deletions as in *can't* → [k ae_n t] are caused by asynchrony between the nasality and tongue closure features; *substitution only*: Incomplete stop closures, as in *legal* → [l iy g_fr ax l], can be described as the substitution of a complete velar closure with a critical closure, resulting in frication (the [g_fr] is a fricated [g]); *both asynchrony and substitution*: *everybody* → [eh r uw ay], which can be described as the substitution of narrow lip and tongue closures for critical or complete ones (accounting for the reduction of the [v], [b], and [dx]) and asynchrony between the tongue and lips in the middle of the word (accounting for [iy] → [uw] via early lip closure)[3]. Livescu and Glass represent this type of model using a DBN and show improved coverage of observed pronunciations and reduced error rate in a lexical access task on the phonetically-transcribed portion of Switchboard (Greenberg *et al.*, 1996), relative to a phone-based pronunciation model.

There are a number of ways in which such a pronunciation model could be incorporated into a complete recognizer. One recent attempt was described in Section V B, in the context of landmark-based speech recognition. A similar pronunciation model has been applied to the task of visual speech recognition (i.e., lipreading) by Saenko *et al.* (2005a,b) and Lee and Wellekens (2001) describe a lexicon using phonetic features.

F. Recognition by articulatory synthesis

Blackburn (Blackburn and Young, 2000) investigated an articulatory speech production model (SPM) in order to give an explicit model of co-articulation. Experiments using real articulatory data were carried out on the Wisconsin X-ray microbeam data (Westbury, 1994) and other experiments on the resource management (RM) corpus (Price *et al.*, 1988). The system rescored output from an HMM recognizer by re-synthesizing articulatory traces from time-aligned phone sequences and mapping these into log-spectra using MLPs (one per phoneme). Errors between these and the original speech were used to reorder the N-best list. The model includes a notion of articulatory effort that leads to an account of the varying strength of coarticulation. On the Wisconsin corpus, recognition performance was enhanced for all but one speaker in the test set using N-best lists with $2 \leq N \leq 5$. The SPM worked best for speakers with low initial word accuracy. On the RM corpus, N-best rescoring for small N offered modest gains, but performance deteriorated with $N = 100$.

VI. DISCUSSION

A. The use of data

This article has given an overview of many approaches to using knowledge about speech production to improve automatic speech recognition. Some require articulatory measurement data, although perhaps only when training the models. Others use a more abstract representation, such as articulatory features or landmarks, which can be obtained more easily.

Approaches that require actual articulatory data will always have to deal with the problems of very limited corpora and with the challenge of doing without this data when actually performing recognition. Elegant solutions to the latter problem include Stephenson's use of DBN variables that are observed during training but hidden during recognition – a technique reminiscent of multi-task training (Caruana, 1997). However, there is still more work to be done, because new (and less invasive) forms of articulatory measurement are becoming available, such as real-time magnetic resonance imaging (Narayanan *et al.*, 2004) or ultrasound. These systems offer great potential because they provide a complete contour of the tongue.

On the other hand, approaches that can utilize *knowledge* about articulation, such as articulatory features that can be initialized from phonetic transcriptions or models with kinematic constraints on (pseudo-)articulator movement, suffer less from the lack of corpora and so are perhaps more likely to transfer easily to larger tasks.

B. Explicit versus implicit modeling

The use of an explicit representation of speech production in the statistical model used for ASR allows the model to make a direct and interpretable account of the processes mentioned earlier. The behavior of such models is more easily analyzed than a large state-tied HMM system and therefore it is, in theory, possible to determine if the model is indeed learning to model specific speech processes.

The price paid for this transparency is typically that the wide variety of powerful techniques developed for HMMs are not immediately available. In some cases, this is merely for practical reasons: for example, algorithms for adaptation or discriminative training are currently more readily available for HMMs than DBNs. In other cases, there are theoretical difficulties: for example, the use of Gaussian *mixture* distributions in LDMs (Section V A 5) leads to intractable models.

A currently underexplored area of research is the marriage of speech production inspiration with standard models such as HMMs, DBNs or ANNs. We have seen some initial work in this area: in Section IV B 1 we described systems which first used ANNs to recover AFs from speech, then used HMMs to model these AFs either by deriving phone class posteriors (this is known as a hybrid HMM/ANN system) or by using the AFs as observa-

tions to be generated by the HMM. This latter method is essentially a Tandem system (Ellis *et al.*, 2001), but without the dimensionality reduction/decorrelation step. A true Tandem system using ANNs trained to recover AFs is a promising area to explore, as shown by Çetin *et al.* (2007), and may be particularly appropriate in a multilingual or language-independent situation. One can argue that it is far easier to devise a universal AF set than a universal phoneme set. So, whilst the explicit use of a speech production representation allows direct modeling of speech effects, implicit approaches like Tandem currently offer a better selection of powerful models and techniques.

C. Moving into the mainstream

There are two distinct routes by which the work we have discussed could move into the mainstream. The first is obvious: if these techniques can show real accuracy gains on the large vocabulary, very large corpus, conversational telephone speech tasks that drive research on conventional HMM-based systems then they may *replace* such systems. The second route is a little more subtle: Speech production-based models can *influence* HMMs-of-phones systems. For example, if it can be shown that a factored state representation provides a more structured or parsimonious state space and therefore allows more sophisticated parameter tying schemes, then this could be used directly in HMM systems, where the factored state is only required during training and can be "flattened" to a single hidden variable so that the model becomes a standard HMM (and can then be used in existing decoders – a major advantage). This transfer of techniques into the mainstream has the added practical advantage that the novel models can continue to be developed on smaller corpora than are currently in use in mainstream HMM research.

D. Ongoing work

In the work that we have mentioned, several strands of research can be identified that continue to be areas of active research. In particular, we wish to highlight DBNs as a very exciting framework (Zweig and Russell, 1998). With the advent of powerful toolkits such as the graphical models toolkit GMTK (Bilmes, 2002) and the Bayes Net Toolbox for Matlab (Murphy, 2001) it is now straightforward to quickly explore a very large family of models. Many of the models mentioned in this article can be implemented in the DBN framework, including all HMMs, the hybrid HMM/BN model in Section V A 4, linear dynamic models, factorial HMMs (Ghahramani and Jordan, 1995) and segmental models. Work in other formalisms continues too. For example, landmark-based systems, as described in section V B are benefiting from the incorporation of classifiers such as SVMs. Indeed, the most successful speech production approaches to ASR generally follow the key principles of conventional techniques: Statistical models are used, parameters are learned from data; these models are used in a consistent probabilistic framework, where evidence from all sources (e.g., the acoustic signal, the lexicon and the language model) are combined to reach the final decision.

E. Evaluation

Over the many decades of development of conventional ASR, a single standard evaluation methodology has emerged: systems are trained and evaluated on standard corpora, and compared using the standard measure of word error rate. Whilst an imperfect measure in some regards, the universal use of WER makes cross-system comparison easy and fair.

For speech production-inspired systems, there is not yet a single evaluation methodology. This is a severe problem both in terms of the development of such methods and their acceptance into the mainstream. Those systems that perform the full ASR task and produce word transcriptions can, of course, be evaluated using WER. However, it is necessary to be able to evaluate systems under development: those that do not (yet) perform word transcription.

The lack of standard evaluation methods hampers development because it is difficult to make cross-system comparisons and thus identify the best approaches. In this paper, we have attempted to make these comparisons wherever possible, but have been limited by the lack of common corpora, task definitions and error measures. Below, we suggest ways in which future comparisons could be made easier.

1. Standard corpora

It is often the case that novel acoustic models cannot be developed on very large corpora (for computational reasons) and it is also often desirable to use relatively simple tasks, such as isolated words or small vocabularies (to make decoding times shorter, or error analysis easier, for example). Typical spontaneous speech corpora have vocabularies that are too large for this purpose. On the other hand, the spontaneous speech effects that production-inspired approaches aim to model are less prominent in read-text corpora (e.g., spoken digits, newspaper sentences). One solution is to construct a small vocabulary corpus from fragments of a large, spontaneous speech corpus, as has been done in the *SVitchboard 1* corpus (King *et al.*, 2005), which contains a number of manageable, yet realistic, benchmark tasks made from Switchboard 1 data.

2. Standard error measures

a. Directly measuring feature recognition accuracy Evaluation of AF recognition accuracy is problematic because comparing recognizer output to reference feature labels derived from phone labels will incorrectly penalize

a number of the processes which the feature models are intended to capture but are not present in the reference transcription. Making comparisons at the frame level will penalize instances where the feature models change value asynchronously. This may be alleviated through the use of a recognition accuracy measure in which timing is ignored, though all feature insertions, deletions and substitutions will still be counted as errors even where they are in fact correct.

Evaluation of landmark accuracy is also problematic since not only are both temporal and classification errors possible, there also is the possibility of insertion or deletion of landmarks. Each researcher currently appears to use a different measure.

Niyogi *et al.* (1999) use receiver operating characteristic (ROC) curves which show the trade off between false detections (landmark insertions) and false rejections (landmark deletions). Automatically detected landmarks are compared to the landmarks in the reference transcription within some time window to allow for small temporal misalignments. Juneja (2004) uses two measures. The first is the frame-level accuracy of a small number of binary manner classifiers. This measure gives a very incomplete picture of the system's performance. The second measure is for the sequence of recovered manner segments and uses the string measure "Percent correct", which does not take into account the large number of insertions that many event-based systems are prone to. "Accuracy" figures are not given. Hasegawa-Johnson *et al.* (2005) feed fragments of speech waveform (half of which contain a reference landmark, and half of which do not) to the landmark detector and the detection accuracy is measured.

b. Evaluating in terms of other linguistic units One option for evaluation is to convert to either phones or syllables, and evaluate using a conventional WER-like measure. This can give some insights into system performance but care must be taken if a fair comparison is to be made. A conversion from AFs to phones at the frame level, as done by King and Taylor (2000), is straightforward, subject to the caveat above that phone-to-feature conversion penalizes some of the very properties of features that are thought to be most desirable.

However, if using a system that incorporates some model of the syllable or word, conversion to phones for evaluation purposes is unfair since the phonotactic constraints of syllables or words provide a strong language model that may not be part of the systems being compared to.

c. Evaluating pronunciation modeling Measures of performance of a pronunciation model include *coverage*, the proportion of spoken pronunciations which the model considers to be allowable realizations of the correct word, and *accuracy*, the proportion of the test set for which the word is recognized correctly. Coverage can be increased trivially, by giving all possible pronunciations of every word a non-zero probability, but this would reduce accuracy by introducing confusability.

F. Future directions

Some powerful classifiers, such as SVMs, are inherently binary (that is, they can only solve two-class problems). In standard ASR systems, such classifiers can only normally be used by reformulating ASR as a two-class problem; for example, disambiguating confusable word pairs from confusion networks (e.g., Layton and Gales, 2004) or in event/landmark-based systems (Hasegawa-Johnson *et al.*, 2005; Juneja, 2004; Juneja and Espy-Wilson, 2003b). Some articulatory feature systems (e.g., SPE, Section II B) are naturally binary, so would be ideal for use with these classifiers.

The phonetic layer in most current systems is a bottleneck. As we have described, it is highly unsatisfactory for describing many forms of pronunciation variation. Some of the feature-based systems we have described still use a phone-based representation between features and word. This clearly constrains the flexibility afforded by the features; for example, it will not allow modeling of highly reduced pronunciations such as the *everybody* → [eh r uw ay] example from Section V E because it prevents modeling asynchronous feature value changes. The problem of mediating between acoustic and word levels, whilst avoiding the phone(me) bottleneck, is addressed from a rather different angle by Gutkin and King (2005) who use a structural approach to discover hierarchy in speech data.

Finally, the potential for language-independent recognition systems based on AFs is huge. This is an almost unexplored area (Stüker, 2003).

G. Suggested background reading

Löfqvist (1997), Perkell (1997) and Farnetani (1997) are all chapters in Hardcastle and Laver (1997), which contains many other interesting articles, such as Steven's chapter on articulatory-acoustic relationships (Stevens, 1997), and a long bibliography. Extensive reading lists for many topics are available from Haskins Laboratories' "Talking Heads" website. For papers on novel approaches to ASR, the proceedings of the Beyond HMM Workshop (2004) are a good starting point. For general background on machine-learning and pattern recognition, we recommend: Bishop and Hinton (1995) and MacKay (2003); for dynamic Bayesian networks either Cowell *et al.* (1999) for the theory, or Bilmes and Bartels (2005) for the use of graphical models in ASR.

Acknowledgments

Frankel, Richmond and Wester are funded by grants from the Engineering and Physical Sciences Research Council (EPSRC), UK and from Scottish Enterprise, UK.

King holds an EPSRC Advanced Research Fellowship. Livescu was funded by an NSF grant and a Luce Postdoctoral Fellowship. Thanks to Mark Hasegawa-Johnson for pointers to various landmark accuracy measures.

Notes

[1]http://sail.usc.edu/span

[2]Caution must be exercised when using such a "N+1"-best list because, if the N hypotheses from the first pass are very poor, then even a bad second-pass model will be able to pick out the correct hypothesis (the "+1") quite easily. Picone *et al.* worked around this problem by also rescoring the list with a standard HMM system for comparison.

[3]This example is taken from the phonetically transcribed subset of the Switchboard database (Greenberg, 1997; Greenberg *et al.*, 1996). All three examples use the ARPABET alphabet modified by diacritics as in the Switchboard phonetic transcriptions.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (**May 1978**), "Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Computer Sorting Technique," The Journal of the Acoustical Society of America **63**(5), 1535–1555.

Bates, R. (**2003**), "Speaker Dynamics as a Source of Pronunciation Variability for Continuous Speech Recognition Models," Ph.D. thesis, Department of Electrical Engineering, University of Washington.

Bilmes, J. (**October 2002**), *GMTK: The Graphical Models Toolkit*, URL `http://ssli.ee.washington.edu/~bilmes/gmtk/`.

Bilmes, J. (**2004**), "What HMMs can't do," in *Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop* (Kyoto, Japan), vol. 104, SP2004-81-95.

Bilmes, J. and Bartels, C. (**September 2005**), "Graphical Model Architectures for Speech Recognition," IEEE Signal Processing Magazine **22**(5), 89–100.

Bilmes, J., Zweig, G., Richardson, T., Sandness, E., Jackson, K., Livescu, K., Xu, P., Holtz, E., Torres, K., J .and Filali, and Brandman, Y. (**2001**), "Discriminatively Structured Graphical Models for Speech Recognition," Johns Hopkins University, CSLP 2001 Summer Workshop Final Report.

Bishop, C. M. and Hinton, G. (**1995**), *Neural Networks for Pattern Recognition* (Clarendon Press, Oxford).

Bitar, N. N. and Espy-Wilson, C. Y. (**May 1995**), "A signal representation of speech based on phonetic features," in *Proc. 1995 IEEE Dual-Use Technologies and Applications Conference* (SUNY Inst. of Tech., Utica/Rome), pp. 310–315.

Bitar, N. N. and Espy-Wilson, C. Y. (**1996**), "A Knowledge-Based Signal Representation for Speech Recognition," in *Proc. ICASSP '96* (Atlanta, Georgia), pp. 29–32.

Blackburn, C. S. and Young, S. J. (**March 2000**), "A self-learning predictive model of articulator movements during speech production," The Journal of the Acoustical Society of America **107**(3), 1659–1670.

Bourlard, H. and Morgan, N. (**1993**), *Connectionist Speech Recognition: A Hybrid Approach* (Kluwer Academic Publishers, Boston).

Bridle, J. S. (**2004**), "Towards Better Understanding of the Model Implied by the use of Dynamic Features in HMMs," in *Proceedings of the International Conference on Spoken Language Processing* (Jeju, Korea), CD-ROM.

Browman, C. and Goldstein, L. (**1991**), "Gestural structures: distinctiveness, phonological processes, and historical change," in *Modularity and the Motor Theory of Speech Perception*, edited by I. Mattingly and M. Studdert-Kennedy (Laurence Erlbaum Associates, Hillsdale, N.J.), chap. 13, pp. 313–338.

Browman, C. and Goldstein, L. (**1992**), "Articulatory phonology: an overview," Phonetica **49**, 155–180.

Caruana, R. (**1997**), "Multitask Learning, Machine Learning," Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Çetin, O., Kantor, A., King, S., Bartels, C., Magimai-Doss, M., Frankel, J., and Livescu, K. (**2007**), "An Articulatory Feature-Based Tandem Approach and Factored Observation Modeling," in *Submitted to Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2007)*.

Chang, J. and Glass, J. (**1997**), "Segmentation and modeling in segment-based recognition," in *Proceedings of Eurospeech* (Rhodes, Greece), CD-ROM.

Chang, S., Wester, M., and Greenberg, S. (**November 2005**), "An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language," Speech Communication **47**(3), 290–311.

Choi, J.-Y. (**1999**), "Detection of Consonant Voicing: A Module for a Hierarchical Speech Recognition System," Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Chomsky, N. and Halle, M. (**1968**), *The Sound Pattern of English* (Harper & Row, New York, NY).

Cohen, J., Kamm, T., and Andreou, A. (**May 1995**), "Vocal tract normalization in speech recognition: compensating for systematic speaker variability," The Journal of the Acoustical Society of America **97**(5), 3246–3247.

Cole, R., Noel, M., Lander, T., and Durham, T. (**1995**), "New telephone speech corpora at CSLU," in *Proceedings of the Fourth European Conference on Speech Communication and Technology*, vol. 1, pp. 821–824.

Cole, R., Stern, R. M., and Lasry, M. J. (**1986**), "Performing Fine Phonetic Distinctions: Templates versus Features," in *Invariance and Variability of Speech Processes*, edited by J. S. Perkell and D. Klatt (Lawrence Erlbaum Assoc., Hillsdale, NJ), chap. 15, pp. 325–345.

Cole, R. A., Rudnicky, A. I., Zue, V. W., and Reddy, R. (**1980**), "Speech as Patterns on Paper," in *Perception and Production of Fluent Speech*, edited by R. Cole (Lawrence Erlbaum Associates, Hillsdale, N.J.), pp. 3–50.

Cowell, R., Dawid, A., Lauritzen, S., and Spiegelhalter, D. (**1999**), *Probabilistic Networks and Expert Systems*, Information Science & Statistics Series (Springer-Verlag New York Inc, New York, NY).

Dalsgaard, P., Andersen, O., and Barry, W. (**1991**), "Multilingual label alignment using acoustic-phonetic features derived by neural-network technique," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-91)*, pp. 197–200.

De Mori, R., Laface, P., and Piccolo, E. (**October 1976**), "Automatic detection and description of syllabic features in continuous speech," IEEE Transactions on Acoustics, Speech, and Signal Processing **24**(5), 365–379.

Deng, L. (**April 1992**), "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," Signal Processing **27**(1), 65–78.

Deng, L. and Ma, J. (**December 2000**), "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," The Journal of the Acoustical Society of America **108**(6), 3036–3048.

Deng, L. and Sun, D. (**1994**a), "Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Adelaide, Australia), vol. I, pp. 45–48.

Deng, L. and Sun, D. X. (**May 1994**b), "A Statistical Approach to Automatic Speech Recognition Using the Atomic Units Constructed From Overlapping Articulatory Features," The Journal of the Acoustical Society of America **95**(5), 2702–2719.

Digalakis, V. (**1992**), "Segment-based stochastic models of spectral dynamics for continuous speech recognition," Ph.D. thesis, Boston University Graduate School, Boston, MA.

Dusan, S. and Deng, L. (**2000**), "Acoustic-to-Articulatory Inversion Using Dynamical and Phonological Constraints," in *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling* (Kloster Seeon, Bavaria, Germany), pp. 237–240.

Eide, E. (**2001**), "Distinctive features for use in an automatic speech recognition system," in *Proceedings of Eurospeech* (Aalborg, Denmark), pp. 1613–1616.

Eide, E., Rohlicek, J., Gish, H., and Mitter, S. (**1993**), "A linguistic feature representation of the speech waveform," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 483–486.

Ellis, D., Singh, R., and Sivadas, S. (**September 2001**), "Tandem Acoustic Modeling in Large-Vocabulary Recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-01)* (Salt Lake City, USA).

Erler, K. and Freeman, G. H. (**October 1996**), "An HMM-based speech recogniser using overlapping articulatory features," The Journal of the Acoustical Society of America **100**(4), 2500–13.

Espy-Wilson, C. Y. and Bitar, N. N. (**1995**), "Speech Parameterization Based on Phonetic Features: application to speech recognition," in *Proceedings of Eurospeech* (Madrid, Spain), pp. 1411–1414.

Farnetani, E. (**1997**), "Coarticulation and connected speech processes," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell Publishers, Oxford), chap. 12, pp. 371–404.

Fosler-Lussier, J. E. (**1999**), "Dynamic Pronunciation Models for Automatic Speech Recognition," PhD dissertation, U. C. Berkeley, Berkeley, CA.

Frankel, J. (**2003**), "Linear dynamic models for automatic speech recognition," Ph.D. thesis, The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK.

Frankel, J. and King, S. (**2001**a), "ASR - Articulatory Speech Recognition," in *Proceedings of Eurospeech* (Aalborg, Denmark), pp. 599–602.

Frankel, J. and King, S. (**2001**b), "Speech recognition in the articulatory domain: investigating an alternative to acoustic HMMs," in *Proceedings of the Workshop on Innovations in Speech Processing* (Stratford-upon-Avon, UK), CD-ROM.

Frankel, J. and King, S. (**2005**), "A Hybrid ANN/DBN Approach to Articulatory Feature Recognition," in *Proceedings of Eurospeech* (Lisbon, Portugal), CD-ROM.

Frankel, J., Richmond, K., King, S., and Taylor, P. (**2000**), "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces," in *Proceedings of the International Conference on Spoken Language Processing* (Beijing, China), CD-ROM.

Frankel, J., Wester, M., and King, S. (**2004**), "Articulatory feature recognition using dynamic Bayesian networks," in *Proceedings of the International Conference on Spoken Language Processing* (Jeju, Korea), CD-ROM.

Fujimura, O. (**1986**), "Relative Invariance of Articulatory Movements: An Iceberg Model," in *Invariance and Variability of Speech Processes*, edited by J. S. Perkell and D. Klatt (Lawrence Erlbaum Assoc., Hillsdale, NJ), chap. 11, pp. 226–242.

Fukuda, T. and Nitta, T. (**2003**a), "Noise-robust ASR by Using Distinctive Phonetic Features Approximated with Logarithmic Normal Distribution of HMM," in *Proceedings of Eurospeech* (Geneva, Switzerland), pp. 2185–2188.

Fukuda, T. and Nitta, T. (**2003**b), "Noise-robust Automatic Speech Recognition Using Orthogonalized Distinctive Phonetic Feature Vectors," in *Proceedings of Eurospeech* (Geneva, Switzerland), pp. 2189–2192.

Fukuda, T., Yamamoto, W., and Nitta, T. (**2003**), "Distinctive phonetic feature extraction for robust speech recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 25–28.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**), "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," National Institute of Standards and Technology, NISTIR 4930.

Ghahramani, Z. and Jordan, M. I. (**1995**), "Factorial Hidden Markov Models," in *Proceedings of the Conference of Advances in Neural Information Processing Systems*, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (MIT Press), vol. 8, pp. 472–478.

Glass, J. (**2003**), "A Probabilistic Framework for Segment-Based Speech Recognition," Computer Speech and Language **17**, 137–152.

Glass, J. R. (**1988**), "Finding acoustic regularities in speech : applications to phonetic recognition," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Goldberg, H. G. and Reddy, D. R. (**1976**), "Feature extraction, segmentation and labelling in the Harpy and Hearsay-II systems," The Journal of the Acoustical Society of America **60**(Supplement Number 1), S11.

Greenberg, S. (**1997**), "The Switchboard Transcription Project," 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series Research Report 24, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.

Greenberg, S., Hollenback, J., and Ellis, D. (**1996**), "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *Proceedings of the International Conference on Spoken Language Processing* (Philadelphia, PA), vol. Supplement, pp. 24–27.

Gutkin, A. and King, S. (**2005**), "Detection of Symbolic Gestural Events in Articulatory Data for Use in Structural Representations of Continuous Speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-05)* (Philadelphia, PA), vol. I, pp. 885–888.

Halberstadt, A. and Glass, J. (**1998**), "Heterogeneous measurements and multiple classifiers for speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 995–998.

Hardcastle, W. J. and Laver, J. (Eds.) (**1997**), *The Handbook of Phonetic Sciences* (Blackwell Publishers, Oxford).

Harrington, J. (**1987**), "Acoustic cues for automatic recognition of English consonants," in *Speech Technology: a survey*, edited by M. A. Jack and J. Laver (Edinburgh University Press, Edinburgh), pp. 19–74.

Harris, J. (**1994**), *English Sound Structure* (Blackwell, Oxford).

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K., and Wang, T. (**2005**), "Landmark-based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop," Tech. rep., Johns Hopkins University.

Hazen, T. J., Hetherington, I. L., Shu, H., and Livescu, K. (**June 2005**), "Pronunciation modeling using a finite-state transducer representation," Speech Communication **46**(2), 189–203.

Hiroya, S. and Honda, M. (**March 2004**), "Estimation of Articulatory Movements from Speech Acoustics Using an HMM-Based Speech Production Model," IEEE Transactions on Speech and Audio Processing **12**(2), 175–185.

Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., and Saltzman, E. (**September 1996**), "Accurate Recovery of Articulator Positions from Acoustics: New Conclusions Based on Human Data," The Journal of the Acoustical Society of America **100**(3), 1819–1834.

Hogden, J., Nix, D., and Valdez, P. (**November 1998**), "An Articulatorily Constrained, Maximum Likelihood Approach to Speech Recognition," Tech. rep., Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

Holmes, W. J. and Russell, M. (**January 1999**), "Probabilistic-trajectory segmental HMMs," Computer Speech and Language

13(1), 3–37.

Honda, K., Hirai, H., and Dang, J. (**1994**), "A physiological model of speech production and the implication of tongue-larynx interaction," in *Proceedings of the International Conference on Spoken Language Processing* (Yokohama, Japan), pp. 175–178.

Hoole, P., Zierdt, A., and Geng, C. (**2003**), "Beyond 2D in articulatory data acquisition and analysis," in *Proceedings of the 15th International Congress of Phonetic Sciences* (Barcelona, Spain), pp. 265–268.

Howitt, A. W. (**September 1999**), "Vowel Landmark Detection," in *Proceedings of Eurospeech* (Budapest, Hungary), vol. 6, pp. 2777–2780.

International Phonetic Association (**1999**), *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet* (Cambridge University Press, Cambridge, UK).

Iso, K. (**1993**), "Speech recognition using dynamical model of speech production," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Minneapolis, MN), vol. 2, pp. 283–286.

Jackson, P., Lo, B.-H., and Russell, M. (**2002**), "Data-driven, nonlinear, formant-to-acoustic mapping for ASR," IEE Electronics Letters **38**(13), 667–669.

Juneja, A. (**2004**), "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. thesis, University of Maryland, College Park, MD.

Juneja, A. and Espy-Wilson, C. (**2003**a), "An event-based acoustic-phonetic approach to speech segmentation and E-set recognition," in *Proceedings of the 15th International Congress of Phonetic Sciences* (Barcelona, Spain), CD-ROM.

Juneja, A. and Espy-Wilson, C. (**2003**b), "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in *Proceedings of the International Joint Conference on Neural Networks* (Portland, Oregon).

Junqua, J. (**January 1993**), "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognisers," The Journal of the Acoustical Society of America **93**(1), 510–524.

Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., and Sen, Z. (**2001**), "What Kind of Pronunciation Variation is Hard for Triphones to Model?" in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 577–580.

Kaburagi, T. and Honda, M. (**May 1996**), "A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes," The Journal of the Acoustical Society of America **99**(5), 3154–3170.

King, S., Bartels, C., and Bilmes, J. (**2005**), "SVitchboard 1: Small Vocabulary Tasks from Switchboard 1," in *Proceedings of Interspeech* (Lisbon, Portugal), CD-ROM.

King, S., Stephenson, T., Isard, S., Taylor, P., and Strachan, A. (**1998**), "Speech recognition via phonetically featured syllables," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 3, pp. 1031–1034.

King, S. and Taylor, P. (**October 2000**), "Detection of Phonological Features in Continuous Speech using Neural Networks," Computer Speech and Language **14**(4), 333–353.

Kirchhoff, K. (**1996**), "Syllable-level desynchronisation of phonetic features for speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 4, pp. 2274–2276.

Kirchhoff, K. (**1999**), "Robust Speech Recognition Using Articulatory Information," Ph.D. thesis, University of Bielefeld, Bielefeld, Germany.

Kirchhoff, K., Fink, G., and Sagerer, G. (**July 2002**), "Combining acoustic and articulatory feature information for robust speech recognition," Speech Communication **37**(3–4), 303–319.

Krstulović, S. (**September 1999**), "LPC-based inversion of the DRM articulatory model," in *Proceedings of Eurospeech* (Budapest, Hungary), vol. 1, pp. 125–128.

Ladefoged, P. (**1997**), "Linguistic Phonetic Descriptions," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell Publishers, Oxford), chap. 19, pp. 581–618.

Layton, M. and Gales, M. (**June 2004**), "Maximum Margin Training of Generative Kernels," Tech. Rep. CUED/F-INFENG/TR.484, Cambridge University Engineering Department.

Lee, K.-T. and Wellekens, C. J. W. (**2001**), "Dynamic lexicon using phonetic features," in *Proceedings of the Eurospeech* (Aalborg, Denmark).

Leonard, R. (**1984**), "A database for speaker-independent digit recognition." in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (San Diego, California, USA), vol. 1, pp. 328–331.

Liberman, A. M. and Mattingly, I. G. (**1985**), "The Motor Theory of Speech Perception Revised," Cognition **1**(36).

Livescu, K. (**September 2005**), "Feature-based pronunciation modeling for automatic speech recognition," Ph.D. thesis, MIT EECS department.

Livescu, K. and Glass, J. (**2004**a), "Feature-based pronunciation modeling for speech recognition," in *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting* (Boston, USA), CD-ROM.

Livescu, K. and Glass, J. (**2004**b), "Feature-based pronunciation modeling with trainable asynchrony probabilities," in *Proceedings of the International Conference on Spoken Language Processing* (Jeju, South Korea), CD-ROM.

Livescu, K., Glass, J., and Bilmes, J. (**2003**), "Hidden feature modeling for speech recognition using dynamic Bayesian networks," in *Proceedings of Eurospeech* (Geneva, Switzerland), vol. 4, pp. 2529–2532.

Lochschmidt, B. (**1982**), "Acoustic-phonetic analysis based on an articulatory model," in *Automatic Speech Analysis and Recognition*, edited by J.-P. Hayton (D. Reidel, Dordrecht), pp. 139–152.

Löfqvist, A. (**1997**), "Theories and models of speech production," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell Publishers, Oxford), chap. 13, pp. 405–426.

Luo, X. and Jelinek, F. (**1998**), "Nonreciprocal Data Sharing in Estimating HMM Parameters," in *Proceedings of the International Conference on Spoken Language Processing*.

Ma, J. and Deng, L. (**2004**a), "A mixed-level switching dynamic system for continuous speech recognition," Computer Speech and Language **18**, 49–65.

Ma, J. and Deng, L. (**January 2004**b), "Target-directed mixture linear dynamic models for spontaneous speech recognition," IEEE Transactions on Speech and Audio Processing **12**(1), 47–58.

MacKay, D. J. C. (**2003**), *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge).

Maddieson, I. (**1997**), "Phonetic Universals," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell Publishers, Oxford), chap. 20, pp. 619–639.

Markov, K., Dang, J., Iizuka, Y., and Nakamura, S. (**2003**), "Hybrid HMM/BN ASR system integrating spectrum and articulatory features," in *Proceedings of Eurospeech*, vol. 2, pp. 965–968.

McAllaster, D., Gillick, L., Scattone, F., and Newman, M. (**1998**), "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *Proceedings of the the International Conference on Spoken Language Processing* (Sydney, Australia), vol. 5, pp. 1847–1850.

McDermott, E. (**2004**), "Production models for speech recognition," in *Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop* (Kyoto, Japan), vol. 104, pp. 1–6, SP2004-81-95.

Metze, F. and Waibel, A. (**2002**), "A Flexible Stream Architecture for ASR Using Articulatory Features," in *Proceedings of the International Conference on Spoken Language Processing* (Denver, CO), CD-ROM.

Murphy, K. (**2001**), "The Bayes Net Toolbox for Matlab," Com-

puting Science and Statistics **33**, CD-ROM.

Muzumdar, M. (**1996**), "Automatic Acoustic Measurement Optimization for Segmental Speech Recognition," Master's thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (**April 2004**), "An approach to real-time magnetic resonance imaging for speech production," Journal of the Acoustical Society of America **115**(4), 1771–1776.

Nix, D. and Hogden, J. (**1998**), "Maximum-Likelihood Continuity Mapping (MALCOM): An Alternative to HMMs," in *Proceedings of the Advances in Neural Information Processing Systems Conference, NIPS*, edited by M. Kearns, S. Solla, and D. Cohn (MIT Press), vol. 11, pp. 744–750.

Niyogi, P., Burges, C., and Ramesh, P. (**1999**), "Distinctive Feature Detection Using Support Vector Machines," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-99)* (Phoenix, AZ).

Nock, H. J. (**2001**), "Techniques for Modelling Phonological Processes in Automatic Speech Recognition," Ph.D. thesis, Cambridge University Engineering Department, Cambridge, UK.

Omar, M. K. and Hasegawa-Johnson, M. (**2002**), "Maximum Mutual Information Based Acoustic Features Representation of Phonological Features for Speech Recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 1–81.

Ostendorf, M. (**1999**), "Moving beyond the 'beads-on-a-string' model of speech," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop* (Keystone, Colorado, USA), vol. 1, pp. 79–83.

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zachs, J., and Levy, S. (**August 1992**), "Inferring Articulation and Recognising Gestures from Acoustics with a Neural Network Trained on X-ray Microbeam Data," The Journal of the Acoustical Society of America **92**(2), 688–700.

Perkell, J. (**1997**), "Articulatory processes," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell Publishers, Oxford), chap. 11, pp. 333–370.

Picone, J., Pike, S., Regan, R., Kamm, T., Bridle, J., Deng, L., Ma, Z., Richards, H., and Schuster, M. (**1999**), "Initial Evaluation of Hidden Dynamic Models on Conversational Speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Phoenix, AZ), vol. 1, pp. 109–112.

Price, P., Fisher, W. M., Bernstein, J., and Pallett, D. S. (**1988**), "The DARPA 1000-word resource management database for continuous speech recognition." in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (New York, NY), vol. 1, pp. 651–654.

Rahim, M., Goodyear, C., Kleijn, W., Schroeter, J., and Sondhi, M. (**February 1993**), "On the Use of Neural Networks in Articulatory Speech Synthesis," The Journal of the Acoustical Society of America **93**(2), 1109–1121.

Richards, H. B. and Bridle, J. S. (**1999**), "The HDM: A Segmental Hidden Dynamic Model of Coarticulation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Phoenix, AZ), vol. 1, pp. 357–360.

Richardson, M., Bilmes, J., and Diorio, C. (**2000**a), "Hidden-Articulator Markov Models for Speech Recognition," in *Proceedings of ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium, ISCA Tutorial and Research Workshop* (Paris, France), pp. 133–139.

Richardson, M., Bilmes, J., and Diorio, C. (**2000**b), "Hidden-Articulator Markov Models: Performance Improvements and Robustness to Noise," in *Proceedings of the International Conference on Spoken Language Processing* (Beijing, China), CD-ROM.

Richmond, K. (**2002**), "Estimating Articulatory Parameters from the Acoustic Speech Signal," Ph.D. thesis, Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK.

Richmond, K., King, S., and Taylor, P. (**2003**), "Modelling the uncertainty in recovering articulation from acoustics," Computer Speech and Language **17**(2), 153–172.

Riley, M. D. and Ljolje, A. (**1996**), "Automatic generation of detailed pronunciation lexicons," in *Automatic Speech and Speaker Recognition: Advanced Topics*, edited by C.-H. Lee, F. K. Soong, and K. K. Paliwal (Kluwer Academic Publishers), pp. 285–302.

Rose, R. C., Schroeter, J., and Sondhi, M. M. (**March 1996**), "The potential role of speech production models in automatic speech recognition," The Journal of the Acoustical Society of America **99**(3), 1699–709.

Rosti, A.-V. I. (**2004**), "Linear Gaussian Models for Speech Recognition," Ph.D. thesis, Cambridge University Engineering Department, Cambridge, UK.

Roweis, S. (**1999**), "Data Driven Production Models for Speech Processing," Ph.D. thesis, California Institute of Technology, Pasadena, California, USA.

Rubin, P. and Vatikiotis-Bateson, E. (**1998**), "Measuring and modeling speech production in humans," in *Animal Acoustic Communication: Recent Technical Advances*, edited by S. L. Hopp and C. S. Evans (Springer-Verlag, New York), pp. 251–290.

Russell, M. and Jackson, P. (**2005**), "A multiple-level linear/linear segmental HMM with a formant-based intermediate layer," Computer Speech and Language **19**(2), 205–225.

Saenko, K., Livescu, K., Glass, J., and Darrell, T. (**2005**a), "Production domain modeling of pronunciation for visual speech recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 473–476.

Saenko, K., Livescu, K., Siracusa, M., Wilson, K., Glass, J., and Darrell, T. (**2005**b), "Visual speech recognition with loosely synchronized feature streams," in *Proceedings of the International Conference on Computer Vision* (Beijing, China), CD-ROM.

Saltzman, E. and Munhall, K. G. (**1989**), "A Dynamical Approach to Gestural PAtterning in Speech Production," Ecological Psychology **4**, 333–382.

Saraclar, M., Nock, H., and Khudanpur, S. (**April 2000**), "Pronunciation modeling by sharing Gaussian densities across phonetic models," Computer Speech and Language **14**(2), 137–160.

Seide, F., Zhou, J., and Deng, L. (**2003**), "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM — MAP decoding and evaluation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Hong Kong, China), vol. 1, pp. 748–751.

Shirai, K. and Kobayashi, T. (**June 1986**), "Estimating Articulatory Motion from Speech Wave," Speech Communication **5**(2), 159–170.

Shu, H. and Hetherington, I. L. (**2002**), "EM training of finite-state transducers and its application to pronunciation modeling," in *Proceedings of the International Conference on Spoken Language Processing* (Denver, CO), CD-ROM.

Stephenson, T., Bourlard, H., Bengio, S., and Morris, A. (**2000**), "Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables," in *Proceedings of the International Conference on Spoken Language Processing* (Beijing, China), vol. 2, pp. 951–954.

Stephenson, T. A., Magimai-Doss, M., and Bourlard, H. (**December 2004**), "Speech recognition with auxiliary information," IEEE Transactions on Speech and Audio Processing **12**(3), 189–203.

Stevens, K. N. (**1997**), "Articulatory-Acoustic-Auditory Relationships," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell Publishers, Oxford), chap. 15, pp. 462–506.

Stevens, K. N. (**2000**), "From acoustic cues to segments, features and words," in *Proceedings of the International Conference on Spoken Language Processing* (Beijing, China), vol. 1, pp. A1–A8.

Stevens, K. N. (**April 2002**), "Toward a model of lexical access based on acoustic landmarks and distinctive features," The Journal of the Acoustical Society of America **111**(4), 1872–91.

Stone, M. (**1997**), "Articulatory processes," in *The Handbook of*

Phonetic Sciences, edited by W. J. Hardcastle and J. Laver (Blackwell Publishers, Oxford), chap. 1, pp. 12–32.

Stüker, S. (**2003**), "Multilingual Articulatory Features," Master's thesis, Carnegie Mellon University, Pittsburgh, PA.

Sun, J., Jing, X., and Deng, L. (**2000**), "Data-driven model construction for continuous speech recognition using overlapping articulatory features," in *Proceedings of the International Conference on Spoken Language Processing* (Beijing, China), CD-ROM.

Tang, M., Seneff, S., and Zue, V. (**2003**), "Two-Stage Continuous Speech Recognition Using Feature-Based Models: A Preliminary Study," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop* (U. S. Virgin Islands), pp. 49–54.

Toda, T., Black, A., and Tokuda, K. (**2004**), "Acoustic-to-Articulatory Inversion Mapping with Gaussian Mixture Model," in *Proceedings of the International Conference on Spoken Language Processing* (Jeju, Korea), CD-ROM.

Tokuda, K., Zen, H., and Kitamura, T. (**2004**), "Reformulating the HMM as a Trajectory Model," in *Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop* (Kyoto, Japan), vol. 104, SP2004-81-95.

Wakita, H. (**1979**), "Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: The State of the Art," IEEE Transactions of Acoustical Speech Signal Processing **ASSP-27**, 281–285.

Weintraub, M., Taussig, K., Hunicke-Smith, K., and Snodgrass, A. (**1996**), "Effect of speaking style on LVCSR performance," in *Proceedings of the International Conference on Spoken Language Processing* (Philadelphia, PA), pp. 16–19.

Westbury, J. (**1994**), *X-Ray Microbeam Speech Production Database User's Handbook*, University of Wisconsin, Madison, WI.

Wester, M. (**2003**), "Pronunciation modeling for ASR – knowledge-based and data-derived methods," Computer Speech and Language **17**, 69–85.

Wester, M., Frankel, J., and King, S. (**2004**), "Asynchronous Articulatory Feature Recognition Using Dynamic Bayesian Networks," in *Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop* (Kyoto, Japan), vol. 104, pp. 37–42, SP2004-81-95.

Wester, M., Greenberg, S., and Chang, S. (**2001**), "A Dutch Treatment of an Elitist Approach to Articulatory-Acoustic Feature Classification," in *Proceedings of Eurospeech* (Aalborg, Denmark), pp. 1729–1732.

Wrench, A. A. (**2001**), "A new resource for production modelling in speech technology," in *Proceedings of the Workshop on Innovations in Speech Processing* (Stratford-upon-Avon, UK), cD-ROM.

Wrench, A. A. and Hardcastle, W. J. (**2000**), "A Multichannel Articulatory Speech Database and its Application for Automatic Speech Recognition," in *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling* (Kloster Seeon, Bavaria, Germany), pp. 305–308.

Wrench, A. A. and Richmond, K. (**2000**), "Continuous Speech Recognition Using Articulatory Data," in *Proceedings of the International Conference on Spoken Language Processing* (Beijing, China), CD-ROM.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**2002**), *The HTK Book (for HTK Version 3.2)*, Cambridge University Engineering Department, Cambridge, UK.

Zhou, J., Seide, F., and Deng, L. (**2003**), "Coarticulation Modeling by Embedding a Target-Directed Hidden Trajectory Model into HMM - Model and Training," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Hong Kong, China), vol. 1, pp. 744–747.

Zlokarnik, I. (**May 1995**), "Adding articulatory features to acoustic features for automatic speech recognition," The Journal of the Acoustical Society of America **97**(5), 3246.

Zue, V. (**November 1985**), "The use of speech knowledge in automatic speech recognition," Proceedings of the IEEE **73**(11), 1602–1615.

Zue, V., Glass, J., Phillips, M., and Seneff, S. (**1989**), "The MIT SUMMIT speech recognition system: a progress report," in *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 126–134.

Zweig, G. and Russell, S. J. (**1998**), "Speech Recognition with Dynamic Bayesian Networks," in *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference,* (Madison, WI), pp. 173–180.