INTEGRATED-MULTILINGUAL SPEECH RECOGNITION USING UNIVERSAL PHONOLOGICAL FEATURES IN A FUNCTIONAL SPEECH PRODUCTION MODEL

Li Deng

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

An outline and general design of an integrated-multilingual speech recognizer is presented, focusing on its key novelty of cross-language portability. This recognizer extends the one described in [5] in that the overlapping features designed originally for American English are improved, generalized, and need only a slight expansion to cover Mandarin/Cantonese Chinese and Canadian French. It also enhances the recognizer of [6] in that the object of dynamic modeling is moved from the observable acoustic domain to the hidden production-affiliated variables defined in the task-dynamic model of speech production [15]. Major components of the recognizer and the related training and recognition algorithms are described.

1. INTRODUCTION

We have in the past several years pursued the development of a comprehensive framework for speech recognition based on statistical and computational models of the phonological and physical processes of speech production [4, 5, 6, 7]. Although these previous efforts have been exclusively limited to American English, our framework is ideally suited for integrated-multilingual speech recognition since the phonological and phonetic uniformities across languages are a natural consequence of the structure of our recognizer. Central to the concept of the integrated-multilingualism is the recognizer's cross-language portability — it enables the recognizer to train only on the first N languages' speech data and to perform recognition directly on any new $(N+1)th$ (target) language with no requirements to re-design and to re-train the recognizer.[1] The integrated-multilingual recognition approach is substantially different from the conventional approach to multilingual speech recognition (rf. [8, 10]), where a large amount of training data specific to the target language had to be collected and be labeled before building and fine-tuning the recognizer for the target language.

At the heart of our integrated-multilingual speech recognition framework is a global and functional model (introduced in [4] first) of the top-down human speech communication process cast firmly in a statistical framework. The global model starts from careful specification of a full set of universal phonological features across languages. This feature-specification process takes into account the cross-language commonality in the possible articulatory, acoustic, and auditory consequences arising from implementation of the features in speech utterances. These top-level phonological features are then passed to control a statistical version of the classic task-dynamic model of speech production [15]. The statistical model in our recognizer has most of its parameters trainable from language-independent speech data. The statistical nature of our task-dynamic model permits computation of likelihoods for arbitrary sequences of acoustic observations, thus enabling speech recognition to perform in a conventional top-down fashion without recourse to direct acoustic-to-articulatory and further inversions which have proved difficult due to the well known non-uniqueness and mismatched degree-of-freedom problems. The purpose of this paper is to present the general structure of the various components of the integrated-multilingual speech recognition framework introduced above.

2. PHONOLOGICAL FEATURES ACROSS LANGUAGES

The new feature-specification system generalizes and expands the system published in [5, 6]. In contrast to the features of [5, 6] which formed hierarchically organized five-tupled bundles after an asynchronous overlapping process and then were mapped directly to acoustics,[2] the current features are made explicitly to associate with the (statistical) control parameters governing dynamic properties of the *tract variables* defined in the model of [15, 12]). The new set of features exploit relations and similarities of feature components across languages, thereby offering opportunities to share observation data among languages and to generalize the observations from source language(s) to a target one in training the integrated-multilingual speech recognizer.

Once a full set of features are specified (for potentially all languages in the world[3]), we need to represent the possible feature sequences with their temporal evolution which are responsible for producing speech utterances cor-

[1] For satisfactory speech recognition performance, the recognizer for a target language may prove necessary to subject to an adaptation process. But the initial recognizer built according to our integrated-multilingual framework will not require speech data from the target language.

[2] This mapping was accomplished via stationary-state HMMs in [5] and via nonstationary-state HMMs in [6].

[3] At the time of this writting, a complete feature specification system for American English, Mandarin and Cantonese Chinese, and for Canadian French has been worked out.

responding to words or word sequences (for any arbitrary language). We have accomplished this by using a set of feature-overlapping rules to construct finite-state automatons whose states are indexed by component features. Improving upon the earlier feature-overlapping rules derived from phonemic transcription [5, 6], the current recognizer also exploits syllable structures and intonation patterns in formulating the rules.

Mandarin Chinese is a syllabic language and syllable is a most natural unit to use for organizing feature overlapping for describing speech utterances. Syllables are countably small, totaling to only 1254 distinct ones (derived from 408 toneless base-syllables). For American English, the syllable count is large but each syllable has a well defined internal structure consisting of onset and rhyme (nucleus plus coda) as its constituents [1, 16]. The feature overlaps within consonant clusters of onset and of coda are rather regular, so are the overlaps between onset and nucleus, and those between nucleus and coda. Our current rule set disallows spreads in Tongue features between onset and coda (i.e. cross nucleus) within a syllable. For Velum and Lips features, the cross-nucleus feature spreads are constrained to be from coda to onset only and not from onset to coda. Feature spreads are permitted, with constraints determined by the prosodic constituent boundaries, between adjacent syllables; i.e. between coda (or nucleus if coda is null) of the preceding syllable and onset (or nucleus if onset is null) of the following syllable. Once a syllable is broken down to its constituents, the size of the constituents becomes countably small and hence they are enumerated exhaustively as we have done in implementing the recognizer.

3. INTERFACE OF OVERLAPPING FEATURES TO "TASK" VARIABLES

In our integrated-multilingual recognizer, each feature is associated with a set of parameters characterizing dynamic properties of the tract variables. We use a subset of the 13 tract variables in the latest version of the task-dynamic model [14], where each tract variable, $\mathbf{z}$, is modeled by a critically damped second order system:

$$\frac{d^2\mathbf{z}(t)}{dt^2} + 2\sqrt{K}\frac{d\mathbf{z}(t)}{dt} + K(\mathbf{z}(t) - Z^0) = 0, \qquad (1)$$

which is characterized by the feature-dependent (normalized) stiffness (K_L, K_F, or K_D, non-random) and by the feature-dependent statistical distribution on the point-attractor (Z^0 as a random variable) of the dynamical system. The form of the distribution is chosen according to the physical properties of the tract variable. In the current implementation, closure-constriction-degree attractors (associated with oral or nasal stop consonants), Z^0_L, Z^0_F, and Z^0_D, are zero or positively valued random variables following an exponential distribution (characterized by parameters β_L, β_F, and β_D). Critical-stricture-degree attractors (associated with fricatives) are strict positively valued random variables following an inverse Gaussian distribution [2] (characterized by parameters λ_F and μ_F or by λ_D and μ_D) with the mode centered (initialized during training) at low, critical constriction-degree values appropriate for generating frications. Open-constriction-degree (vocalics mainly) attractors are also strict positively valued inverse-Gaussian random variables with the mode at the constriction sizes appreciably greater than the critical ones. The remaining constriction-degree attractors (Z^0_V, Z^0_X) and all the constriction-location attractors ($Z^0_{\tilde{L}}$, $Z^0_{\tilde{F}}$, $Z^0_{\tilde{D}}$) are again inverse-Gaussian random variables, whose respective distribution parameters λ and μ are trained with initial values set at the nominal ones according to speech production data or principles.

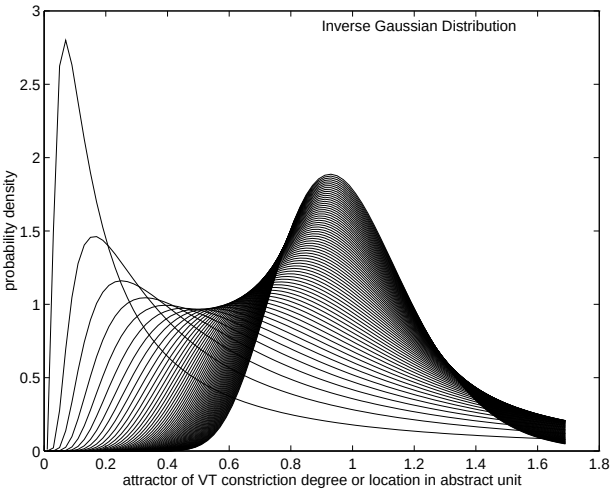

Figure 1. Inverse Gaussian distribution for the attractor of vocal tract constriction degree or constriction location in arbitrary unit.

The probability distribution function (pdf) of the inverse Gaussian distribution is

$$f(x;\mu,\lambda) = \sqrt{\lambda/2\pi} \times exp[-\lambda(x-\mu)^2/2\mu^2 x], \qquad x > 0$$

with μ as the mean of the distribution and λ as the scale (skewness) factor. The choice of this distribution for describing the statistical behavior for the attractors in most of the tract variables is mainly motivated by its flexibility to represent the possible ranges of vocal-tract constriction degrees and locations (which are all positively valued).[4] Figure 1 shows the pdf of the inverse Gaussian distribution for a fixed parameter $\mu = 1$ and with a varying parameter λ from 0.2 to 20 (0.3 as increment). For small values of λ, the pdf's are highly skewed with the mode moving towards zero. When this distribution is used for oral constriction-degree attractors, then fricatives will be automatically trained to associate with small values of λ (thereby the mode becomes close to zero), approximants with the mode away from zero, and vowels with the mode further away from zero. The attractor of the glottal-aperture (velic-aperture) variable will also have the mode in the inverse Gaussian pdf far away from zero for aspiration (nasal) sounds, while that for voiced (non-nasal) sounds will have the mode close to zero.

[4]It is also motivated by well established results [2] for this distribution in parameter estimation, Bayesian inference, significance test, and in regression analysis (both linear and nonlinear), all of which are important in our development of training and recognition algorithms used for the recognizer.

One significant advantage that arises from interfacing the phonological features to the task-dynamic model is that a mechanism similar to that of gesture blending or "parameter tuning" described in [15] can be developed to merge the tract-variable attractors' distributions associated with overlapping features (correlated with "co-produced" articulatory gestures) into a single distribution defined on the same tract-variable coordinate system.[5] Therefore, the technique for Cartesian-product construction of finite-state automata [5] is no longer needed. The total number of primitive (inverse-Gaussian) distributions for characterizing the relationship between the symbolic features and the tract variables for entire American English is as small as forty[6]. This number is increased only up to forty five when French is added, and to only fifty some when Mandarin Chinese is further added into the language pool.

4. FROM TASK VARIABLES TO ACOUSTICS VIA MODEL-ARTICULATORS

Given the time varying tract variables produced from the task-dynamic model, motions of a set of biomechanical model-articulators,[7] $\mathbf{x}$, can be generated via a highly coupled nonlinear kinematic relationship:

$$\mathcal{X} \rightarrow \mathcal{Z}: \quad \mathbf{z}(\mathbf{t}) = \mathcal{Z}(\mathbf{x}(\mathbf{t})). \tag{2}$$

Combining Eqns.(1) and (2) leads to the following dynamic system for model-articulators:

$$J(\mathbf{x})\frac{d^2\mathbf{x}(t)}{dt^2}+[\frac{dJ(\mathbf{x})}{dt}+2\sqrt{K}J(\mathbf{x})]\frac{d\mathbf{x}(t)}{dt}+K[\mathcal{Z}(\mathbf{x}(\mathbf{t}))-Z^0]=0, \tag{3}$$

where $J(\mathbf{x})$ is the Jacobian transformation matrix for Eqn.(2), and $\frac{dJ(\mathbf{x})}{dt}$ is the matrix obtained by differentiating each element of $J(\mathbf{x})$ with respect to time.

Note that the mapping of Eqn.(2) is geometrical in nature,[8] and it also reflects speaker and speaking-mode variabilities (including varying dialects, foreign accents, and speaking rates) in articulation for implementing a given "task" of vocal-tract constriction. For implementation feasibility, we use radial basis function (RBF) neural nets as a device for data interpolation in multi-dimensional space to approximate the mapping in Eqn.(2):

$$\mathbf{z} = \mathcal{Z}(\mathbf{x}) \approx \sum_i w_i e^{(\mathbf{x}-\mu_\mathbf{i})^T r_{P_i}(\mathbf{x}-\mu_\mathbf{i})}.$$

The parameters in the above RBF approximation are initialized during training based on simulations of a geometric articulatory model with a standard vocal tract. The degree-of-freedom problem (one-to-many relation between $\mathbf{z}$ and $\mathbf{x}$) can be addressed by incorporating constraints using techniques similar to "transformation gating" [15] during the RBF network learning phase. Nonsupervised training can be used for multiple sets of RBF parameters to cluster dialect and foreign-accent variabilities.

Given the time varying model-articulator motions produced from Eqn.(3), the observable acoustic signal $\mathbf{O}$ is generated from a further nonlinear mapping:

$$\mathcal{X} \rightarrow \mathcal{O}: \quad \mathbf{O} = \mathcal{O}(\mathbf{x}). \tag{4}$$

This lowest-level mapping is independent of linguistic, dialectic, and speaking-mode factors and is only a function of details of the vocal-tract's acoustic properties such as the total vocal-tract length, the pharyngeal height, the shape of nasal cavity, and the average loss in the vocal-tract's wall vibration, etc.. We use another set of trainable RBFs to approximate this articulator-to-acoustics mapping. The RBF parameters are initialized based on simulations of vocal-tract acoustics using the area functions derived again with a geometric articulatory model.

5. TRAINING/DECODING ALGORITHMS

With the interface between the phonological features and the tract variables, and using the feature blending rules together with the nonlinear one-to-many mapping Eqn.(2) that couples each tract variable to a number of model-articulators, Cartesian-product construction is no longer required in linking the features to acoustics. This has drastically reduced the overall size of the trainable recognizer parameters. The entire parameter set of our recognizer, potentially capable of language-independent, speaker-independent, speaking-style-independent, and unlimited-vocabulary speech recognition, is on the order of two thousands only (about three orders of magnitude lower than that required by the conventional HMM recognizers which do not exploit the internal structure of speech). The model parameters of our recognizer include feature-dependent stiffness ($K's$), feature-dependent attractor distribution parameters ($\lambda's, \mu's, \beta's$), feature blending weights, feature-independent (but dialect/accent-dependent) RBF weights, and, finally, feature-independent (but vocal-tract-size dependent) RBF weights.

The training of the recognizer is accomplished by gradient-descent-based numerical optimization techniques. Gradients of the objective function with respect to the recognizer parameters are computed in an analytic form and are then used for optimization. Due to the relatively small size of the parameter set, this approach is feasible, although it is still very slow. Alternative techniques, such as genetic algorithms [12] or EM-based optimal filtering algorithms

[5] In the recognizer implementation, such merge is accomplished by a linear combination of the attractor random variables. Thus the final resulting distribution becomes numerical convolution of the individual distributions associated with each of the overlapped features.

[6] The number of primitive distributions is on the order of one to three thousands in the recognizer described in [5], and on the order of millions in conventional HMM speech recognizer [10].

[7] In the latest version of the task-dynamic model [14], the model-articulators are expanded from the older version [15] and have included: upper and lower lips, jaw, tongue body, tongue tip, velum, glottal width, total lung force, supralaryngeal vocal tract volume, and vocal fold tension.

[8] To be more precise, the elements in $J(\mathbf{x})$ and $\frac{dJ(\mathbf{x})}{dt}$, which are determined from Eqn.(2), characterize the geometrical relationships between motions of the model-articulators and of their corresponding tract variable.

[13], may prove more effective in the future despite their high implementation complexity.

The recognition algorithm is formulated as a straightforward top-down search problem within the well established Bayesian framework consistent with the mainstream speech recognition approach. No bottom-up inversion from acoustics to articulation (and further to tract variables, phonological features, and to word sequences) is required for recognition. Such problematic inversion is avoided because the statistical formulation of our detailed "forward" speech generation model allows likelihood evaluation for the observable speech acoustics given an arbitrary word sequence.

We should emphasize here that the cross-language portability as a key trait of our recognizer originates from the structure of the recognizer, rather than from the more or less conventional training and recognition algorithms described here. As an example, once we trained the distribution parameters of the tract variables associated with Lips-feature of /u/ and those with TongueDorsum-feature /i/ using *English utterances only*, these distributions will be gated, at the lower model-articulator and acoustic levels, via the nonlinear mappings (Eqns.(2) and (4)) to automatically produce the appropriate distribution for French /y/ and Chinese /y/ *without use of French or Chinese utterances*. Hence, it is the carefully structured components of the recognizer which give the cross-language portability.

6. SUMMARY AND DISCUSSIONS

This paper has provided some accounts of a new speech recognizer, currently under development, which aims at *integrated* multilingualism enjoying cross-language portability. It will potentially overcome many serious limitations of the current, mainstream data-driven approach to speech recognition. One immediate limitation is the large efforts and resources required to perform data collection/labeling and system tuning when a recognizer is ported from one language to another (and from one recognition task to another even within the same language). The drawbacks of the conventional data-driven approach to multilingual speech recognition root in its (intentional but justifiable) ignorance, in the recognizer design, of the internal phonological (symbolic) and phonetic (numeric and dynamic) structures underlying all members of human languages.

The framework described in this paper is a significant extension of that described in [5], where the overlapping features designed originally for American English are modified, generalized, and only slightly expanded to cover a number of other languages. It is also a natural extension of the recognizer described in [6], where the modeling component for the dynamic pattern in speech production is pushed from the surface acoustic domain inwardly to the internal, abstract "task" space spanning the coordinates of the tract variables. Conspicuously missing, however, in the current framework are direct modeling of dynamics on biomechanic articulators and the possibility of using acoustic/perceptual criteria as direct feature correlates.[9] In an attempt to work out a consistent recognition framework unifying potentially all languages, significant difficulties have been encountered with use of acoustic/perceptual criteria in defining "tasks" of speech production for multiple languages. Use of vocal-tract constrictions in a stylized (and normalized) vocal tract to define such tasks, on the other hand, enables effective exploitation of a rich source of speech knowledge across languages (e.g. [3, 9, 11]) and gives a head start in formulating the speech recognition framework. Armed with a powerful statistical formalism, we are confident that the recognizer described in this paper will compensate for any incomplete nature of such knowledge and make a workable system enjoying the desirable integrated-multilingualism.

[9] Both of these two aspects have been included in a separate speech production model described elsewhere [7, 13].

REFERENCES

[1] J. Blevins. "The syllable in phonological theory," in *The Handbook of Phonological Theory,* J. Goldsmith (ed.), Blackwell, Cambridge, 1995, pp. 206-244.

[2] R. Chhikara and J. Folks. *The Inverse Gaussian Distribution – Theory, Methodology, and Applications,* Marcel Dekker, Inc, New York, 1989.

[3] P. Delattre. *Comparing the Phonetic Features of English, French, German, and Spanish,* London: George Harrap & Company, 1965, 118 pages.

[4] L. Deng. "Design of a feature-based speech recognizer," *J. Acoust. Soc. Am.*, vol. 93(4) Pt.2, 1993, pp. 2318.

[5] L. Deng and D. Sun. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Am.*, Vol. 95, No. 5, May 1994, pp. 2702–2719.

[6] L. Deng and H. Sameti. "Transitional speech units and their representation by the regressive Markov states," *IEEE Trans. Speech Audio Proc.*, Vol.4, 1996, pp. 301-306.

[7] L. Deng, G. Ramsay, and D. Sun. "Production models as a structural basis for automatic speech recognition," *Speech Communication* (invited), to appear, July 1997.

[8] J. Glass et al. "Multilingual spoken-language understanding in the MIT Voyager system", *Speech Communication*, Vol. 17, 1995, pp. 1-18.

[9] P. Ladefoged and I. Maddieson. *The Sounds of the World's Languages,* Oxford: Blackwell Publishers, 1996, 425 pages.

[10] L. Lamel and J. L. Gauvain. "Issues in large vocabulary, multilingual speech recognition," *Proc. Eurospeech,* 1995, pp. 185-188.

[11] I. Maddieson. *Patterns of Sounds,* Cambridge: Cambridge University Press, 1984, 420 pages.

[12] R. McGowan. "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," *Speech Communication*, vol. 14, 1994, pp. 19-48.

[13] G. Ramsay and L. Deng. "Optimal filtering and smoothing for speech recognition using a stochastic target model," *Proc. ICSLP*, Vol. 2, 1996, pp. 1113-1116.

[14] P. Rubin et al. "CASY and extensions to the task-dynamic model," *Proc. 4th European Speech Production Workshop*, Autrans, France, 1996, pp. 125-128.

[15] E. Saltzman and K. Munhall. "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, Vol. 1, pp. 333-382, 1989.

[16] E. Selkirk. "The Syllable", in *The Structure of Phonological Representation,* H. Hulst and N. Smith (eds.), Foris Publications, 1982, pp. 337-383.