Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 79 (2016) 61–73

www.elsevier.com/locate/specom

Acoustic-articulatory relationships and inversion in sum-product and deep-belief networks

Frank Rudzicz[a,b,*], Arvid Frydenlund[b], Sean Robertson[b], Patricia Thaine[b]

[a] *Toronto Rehabilitation Institute-UHN, Toronto, Canada*
[b] *University of Toronto, Department of Computer Science, Toronto, Canada*

Received 3 August 2015; received in revised form 4 February 2016; accepted 1 March 2016
Available online 21 March 2016

Abstract

We provide the first direct comparison of sum-product networks (SPNs) and deep-belief networks on speech, and the first application of SPNs to acoustic-articulatory inversion. Interestingly, speech from individuals with cerebral palsy is reconstructed significantly more accurately across all manners of articulation using SPNs than when using DBNs. In order to select appropriate input parameters, we first compare MFCCs, wavelets, scattering coefficients, and vocal 'tract variables' as predictor variables to phonological features. Here, MFCCs provide for more accurate classification over a broad array of phonological categories (in the high 90s in many cases) than the other feature types. All experiments use the MOCHA-TIMIT and TORGO acoustic-articulatory databases.

Keywords: Speech articulation; Wavelets; Scattering coefficients; Sum-product networks.

1. Introduction

A classic perspective of speech production considers each phoneme to be a product of several phonological features which loosely correspond to the positions of articulators (Archibald and O'Grady, 2008). Theoretically, a phonological feature vector (e.g., $[+plosive, -voicing, +bilabial, \ldots]$ for */p/*) uniquely identifies each phoneme, and correct phone classification should inform us of the position of articulators, and vice versa. Several studies have in fact used phonological (or "articulatory") features to classify speech (Frankel et al., 2007; Fukuda et al., 2003), with a particular benefit in the presence of extreme environmental noise (King et al., 2007). For instance, incorporating such features with maximum mutual information into hidden Markov systems can reduce word-error rates from 25% to 19.8% on English spontaneous scheduling tasks (Metze, 2007).

Despite their empirical success, modeling speech articulation discretely cannot inherently account for more complex aspects of articulatory organization for which parallel and self-organizing theories may be more appropriate (Rudzicz et al., 2008; Smith and Goffman, 2004). In order to study the long-term dynamics of speech, we require a framework of dynamical systems into which continuous data can be explored. *Task dynamics* (Saltzman and Munhall, 1989) introduces the notion that the dynamic patterns in speech are caused by overlapping *gestures*, which are high-level abstractions of goal-oriented reconfigurations of the vocal tract, such as bilabial closure, or velar opening. Here, all implicit spatiotemporal behavior underlying speech is the result of the interaction between the abstract *intergestural* dimension (between tasks) and the geometric *interarticulator* dimension (between physical actuators) (Saltzman and Munhall, 1989).

Each gesture occurs within a *tract variable* (TV): lip aperture and protrusion (*LA, LP*), tongue tip constriction location and degree (*TTCL, TTCD*), tongue dorsum constriction location and degree (*TDCL, TDCD*), velar opening (*VEL*), glottal vibration (*GLO*), and lower tooth height (*LTH*). A gesture to close the lips, e.g., would set LA close to zero. Not all of these canonical TVs are used in our current work. The dynamic influence of gestures on the relevant TV is modeled by a non-homogonous second-order linear differential equation

* Corresponding author at: Toronto Rehabilitation Institute-UHN, Toronto, Canada. Tel.: +1 416 597 3422x7971.
E-mail address: frank@cs.toronto.edu (F. Rudzicz).

http://dx.doi.org/10.1016/j.specom.2016.03.001
0167-6393/

analogous to a damped mass-spring (Reimer and Rudzicz, 2010). Here, TVs are derived from electromagnetic articulography (EMA), which tracks the positions of points on the articulators, as described in Section 3.3.

Phonological features and TVs, in a sense, measure the same phenomena. Phonological features are discrete representations of articulatory goals that are easily interpretable but can lead to overgeneralization and quantization error. Tract variables provide a continuous representation of intent subject to (and filtered by) individual articulatory constraints that are relatively detailed but difficult to manipulate. Both approaches represent relevant speech goals and both make assertions about the general position of articulators.

Relating acoustics and articulation through statistical models can be challenging when these modalities are disrupted biologically, as in neuro-motor disorders such as cerebral palsy (CP). In CP, damage to the cranial nerves that control the articulatory musculature of speech (Moore and Dalley, 2005) can result in reduced control of phonation, hypernasality, heavily slurred speech, and a more diffuse and less differentiable vowel target space (Kent and Rosen, 2004). Considering this population, and individuals with *dysarthria* generally, provides an important measure of the robustness of our acoustic-articulatory models.

We design two experiments that relate the phonological and task dynamics theories of speech production to their acoustics, within the context of atypical speech articulation. Experiment 1 (Section 4) broadly examines the relationships between articulatory configurations and three types of acoustic features that differ in their representation of the spectrum, and correlational and discriminative relationships are sought through the use of an SVM classifier. We choose to examine Mel-frequency cepstral coefficients, discrete wavelet coefficients, and scattering coefficients due to the relative ubiquity of the first, and the theoretical advantages of the others, as described in Section 2. Experiment 2 (Section 5) uses the results of that feature analysis to perform tractable acoustic-to-articulatory inversion using deep-belief networks (DBNs) and sum-product networks (SPNs), which provides the first comparison of these two methods for this task. DBNs have already been applied to acoustic-articulatory inversion, but SPNs have several uniquely interesting aspects, including a partition function that is guaranteed to be tractable given certain limitations of the network structure.

2. Acoustic features

In the following experiments, Mel-frequency cepstral coefficients (MFCCs), discrete wavelet coefficients (DWCs), and scattering coefficients (SCs) are considered. Given the relative ubiquity of MFCCs (Ganchev, 2011), only the latter two are discussed here.

2.1. Discrete wavelet coefficients

Wavelets are signal filters localized in time and frequency which can represent transient events in a signal, which can be common in cerebral palsy. A wavelet *transform* is the projection of a signal onto a wavelet. By using the 'dual form' of a collection of wavelets, the original signal can be reconstructed in a numerically stable way, which implies complete characterization of the spectrum (Daubechies, 1992). A wavelet is admissible if it is centered around 0 in the time domain, and has a quick decay (most of its energy surrounds some points $\pm\ \xi_0$). Dilating and translating this 'mother wavelet' allows for the creation of a family of wavelets, which can characterize a signal. Formally, given an admissible mother wavelet ψ, a family of wavelets is

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right)$$

for all time shifts τ and scaling factors a (Quatieri, 2008). A family of wavelets has frame bounds over a signal f, defined in the discrete case as

$$A\|f\|^2 \leq \sum_{m,n} |\langle f, \psi_{m,n}\rangle|^2 \leq B\|f\|^2$$

For $m, n \in \mathbb{Z}$, $0 < A \leq B < \infty$. The family of transforms essentially conserves the energy of the original signal f. In our work, the dyadic wavelet transform (Daubechies, 1992; Mallat, 1989), influenced by multiresolution analysis, computes a wavelet transform with a simple iterative algorithm.

2.2. Scattering coefficients

Andén and Mallat (2014) drew parallels between Mel-frequency spectrum coefficients (MFSCs, i.e., MFCCs prior to the cosine transform) and wavelet transforms. An MFSC is approximately $M(n, l) \approx |f{*}v_l|^2{*}|w|^2(n)$, where v_l is the impulse response of the Mel-scale filter and w is the windowing filter used in calculating short-time Fourier transforms (Mallat, 2012). Intuitively, the MFSC is the convolution of the energy of the Mel-scale subband and a lowpass filter (w typically has a much smaller bandwidth than the equivalent Mel-scale filter). Taking the power of a convolved signal in the time domain effectively transforms it into a low frequency characterization. Therefore v_l captures high frequency components, which are then modulated to the origin, and cleared of noise with w. Intuitively, scattering coefficients (SCs) extend MFCCs by computing modulation coefficients of multiple orders through cascades of wavelet convolutions. They have recently been shown to provide state-of-the-art phone classification with TIMIT (Andén and Mallat, 2014). While MFCCs describe local timescales efficiently at around 16 ms to 25 ms in speech, SCs can serve as stable and invariant signal representations over much larger timescales.

Scattering transforms iteratively apply convolutions and modulations to transform (and recover) high-frequency components into lower bandwidths, then apply a lowpass filter to average the lower bandwidths. The combination of power spectrum modulation and wavelet transforms is a characterization which retains high frequency information while remaining stable to deformation in those frequencies (Andén and Mallat, 2014).