From Acoustics to Vocal-Tract time Functions

Vikramjit Mitra[1], İ Yücel Özbek[3], Hosung Nam[2], Xinhui Zhou[1], Carol Y. Espy-Wilson[1]

[1]Institute for Systems Research, University of Maryland, College Park, [2]Haskins Laboratories, New Haven, [3]Middle East Technical University, Turkey

Email: [1]{vmitra@umd.edu, zxinhui@umd.edu, espy@umd.edu}, [2]{nam@haskins.yale.edu}, [3]{iozbek@illinois.edu}

Introduction

The goal of our research is to develop a gesture and landmark-based speech recognition system. This work presents the initial step to achieve such a system, where the mapping between the speech signal and the **vocal tract time functions** (**VTTF**) is considered.

- VTTFs are time-varying physical realizations of articulatory gestures at distinct vocal tract sites for a given utterance.
- VTTFs describe the geometric features of the vocal tract shape in terms of constriction degree and location.
- VTTFs would help to obtain gestural information from the acoustics.

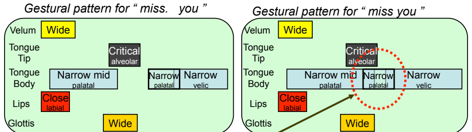
The proposed mapping is based on a hierarchical support vector regression (**SVR**) followed by **Kalman** smoothing. The smoothed VTTFs will be used to recover gestural information from the speech signal.

Motivation

- Automatic Speech Recognition (ASR) suffers from poor performance in casual speech due to acoustic variations
- Phone-based ASR systems suffer due to co-articulation.
  - phone units are distinctive in the cognitive domain but are not invariant in the physical domain.
  - phone-based ASR systems do not adequately model the temporal overlap that occurs in more casual speech.
- To address co-articulation, diphone and triphone models are used.
  - they limit contextual influence to only immediately close neighbors.
  - require a large training data to combinatorially generate all possible diphone or triphone units
- **Articulatory phonology** proposes the articulatory constriction gesture as an invariant action unit and argues that human speech can be decomposed into a constellation of articulatory gestures [1, 2]
  - ✓ This representation allows for temporal overlap between neighboring gestures.
  - ✓ Acoustic variations are accounted for by gestural co-articulation and reduction

How can **Gestures** address ***speech variability***?

Gestural pattern for " miss. you " *Gestural pattern for " miss you "*



The overlap of the tongue gestures for /s/ in "miss" and the /y/ in "you" will change the fricative acoustics. However, at the articulatory level, all of the gestures are there, only the timing and degree of the gestures have changed.

- ❑ **Gestures** can be defined in eight vocal tract (VT) constriction variables shown in Table 1.
- ❑ The activation onset/offset times, and dynamic parameter specifications of constriction gestures and inter-gestural timing patterns are represented by a **Gestural Score**, which is distinct for a given lexical item.

The gestural score for an **active gesture** is specified by the following:

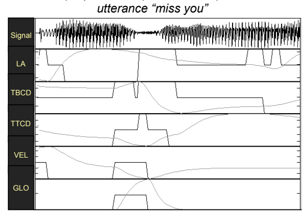
- **Target** defines the constriction location/degree for that particular tract variable on which that gesture is defined.
- **Stiffness** represents the elasticity of a gesture and is proportional to time to achieve the target.
- **Blending** defines how two overlapping gestures corresponding to the same tract variable should be blended with one another

- Gestures can temporally overlap within and across tract variables.
- Even when a tract variable does not have an active gesture, it can be varied passively by another tract variable sharing a common articulator.

Table 1. Constriction organ, vocal tract variables & involved model articulators

Constriction organ	VT variables	Articulators
Lip	Lip Aperture (LA)	Upper lip, lower lip, jaw
	Lip Protrusion (LP)	
Tongue Tip	Tongue tip constriction degree (TTCD)	Tongue body, tip, jaw
	Tongue tip constriction location (TTCL)	
Tongue Body	Tongue body constriction degree (TBCD)	Tongue body, jaw
	Tongue body constriction location (TBCL)	
Velum	Velum (VEL)	Velum
Glottis	Glottis (GLO)	Glottis

Gestural activations (step functions) and VTTFs (smoother curves) for the utterance "miss you"



Gestural activations (the step curve in the above figure) can have three possible values: 0 → Inactive gesture, 1 → active gesture without any intra-gestural blending and 2 → active gesture with intra-gestural blending

The Data

- ❑ Synthetic data was used in this research as no real-world speech database exists for which the ground-truth VTTFs are known (we are in the process of creating ground-truth VTTFs and Gestural scores for X-ray Microbeam).
- ❑ Given English text or ARPABET, **TADA** [4] (TAsk Dynamics Application model) generates input in the form of formants and VTTFs for **HLsyn**™ (quasi-articulator synthesizer, Sensimetrics Inc.).
- ❑ TADA output files were then fed to HLsyn™ to generate acoustic waveform.
- ❑ A synthetic acoustic dataset for 363 words (chosen from Wisconsin X-ray microbeam data) were created
  - VTTFs (sampled at 200Hz) were created by TADA
- ❑ Speech signal is converted to acoustic parameters (APs) [5] (e.g. formant information, mean Hilbert envelope, energy onsets and offsets, periodic and aperiodic energy in subbands [6] etc.)
  - APs were measured at a frequency of 200Hz.
  - 53 APs were considered
  - A subset of the APs was selected for each VTTF based upon their relevance

Hierarchical SVR structure

- Certain VTTFs (TTCL, TBCL, TTCD and TBCD) are known to be functionally dependent upon other VTTFs

$$f_{TTCL} : TTCL \leftarrow (AP_{TTCL}, LA)$$
$$f_{TBCL} : TBCL \leftarrow (AP_{TBCL}, LA)$$
$$f_{TTCD} : TTCD \leftarrow (AP_{TTCD}, TTCL, TBCL, LA)$$
$$f_{TBCD} : TBCD \leftarrow (AP_{TBCD}, TBCL, LA)$$

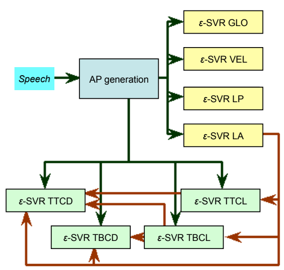
where AP_{XYZ} denotes the set of pertinent APs for VTTF XYZ

- The remaining VTTFs (GLO, VEL, LA and LP) are relatively independent and can be obtained directly from the APs.

The ε-SVR [20] (a generalization of the SVM) works for only single output.

- ❑ ε-SVR uses the parameter ε (the unsusceptible coefficient) to control the number of support vectors.
- ❑ SVR projects input data into a high dimensional space via non-linear mapping and performs linear regression in that space.
- ❑ For 8 VTTFs, 8 different ε-SVRs were created, the hierarchical structure was based upon the correlation amongst the VTTFs.

The hierarchical ε-SVR architecture for generating VTTFs



Mapping from the acoustic domain to the articulatory domain suffers from non-uniqueness (one-to-many mapping).

- ❑ This problem could be ameliorated by incorporating dynamic information at the input space (i.e., using contextual information)

A context window of ***N*** (varied from 5 to 9) is considered for the input:

- ❑ which means ***N*** frames were selected before and after the current frame with a frame shift of 2 (time shift of 10 ms)
- ❑ Input vector has dimension (2***N***+1)****d***, where ***d*** is the dimension of the AP for a particular VTTF

Number of APs, Optimal context window and input dimension for each VTTF

VTTF	# of APs	Optimal *N*	Input Dimension (d)
GLO	15	6	195
VEL	20	7	300
LP	15	6	195
LA	23	8	391
TTCL	22	7	345
TTCD	22	5	275
TBCL	18	5	209
TBCD	18	6	260

***Note**: optimal N is the context window which gives the least MSE*

Post-processing

The estimated VTTFs were found to be noisy so that we used a smoother to help reduce the root-mean-square error (rmse)

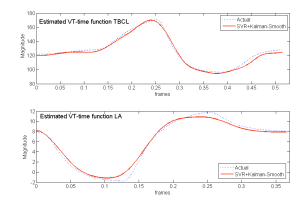
- Two types of smoothing explored

(1) Running average and (2) Kalman smoothing

***rmse** for the different VTTFs*

VTTF	rmse		
	ε-SVR	After averaging filter	After kalman smoothing
GLO	0.039	0.040	0.036
VEL	0.025	0.025	0.023
LP	0.565	0.536	0.508
LA	2.361	2.227	2.178
TTCD	3.537	3.345	3.253
TBCD	1.876	1.749	1.681
TTCL	8.372	8.037	7.495
TBCL	14.292	13.243	12.751

Overlaying plot of the actual VTTF (TBCL & LA) reconstructed VTTF after Kalman smoothing



Conclusion

- Proposed a hierarchical SVR for VTTF estimation from acoustics.
- **Kalman** smoothing helped to reduce rmse by 9.44%.
- Contextual information helped to reduce reconstruction error.

Future Directions

- o Explore other machine learning approaches to perform the same task.
- o Address the issue of non-uniqueness in speech inversion in a probabilistic manner.
- o Explore other acoustic features.
- o Evaluate the system performance when speech is corrupted with noise.

References

[1] C. Browman and L. Goldstein, "Articulatory Gestures as Phonological Units", Phonology, 6: 201-251, 1989
[2] C. Browman and L. Goldstein, "Articulatory Phonology: An Overview", Phonetica, 49: 155-180, 1992
[3] K. Kirchhoff, "Robust Speech Recognition Using Articulatory Information", PhD Thesis, University of Bielefeld, 1999.
[4] H. Nam, L. Goldstein, E. Saltzman and D. Byrd, "Tada: An enhanced, portable task dynamics model in matlab", Journal of the Acoustical Society of America, Vol. 115, Iss. 5, pp. 2430, 2004.
[5] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks", PhD thesis, University of Maryland College Park, December 2004.
[6] O. Deshmukh, C. Espy-Wilson, A. Salomon and J. Singh, "Use of Temporal Information: Detection of the Periodicity and Aperiodicity Profile of Speech", IEEE Trans. on Speech and Audio Processing, Vol. 13(5), pp. 776-786, 2005.

Acknowledgement

This research was supported by NSF grant IIS0703859 and NIH grant DC-02717. We also would like to acknowledge the helpful advice from Elliot Saltzman (BU), Mark Hasegawa-Johnson (UIUC) and Louis Goldstein (USC).