ICASSP 2010, Dallas TX

AN EXPLORATORY STUDY OF MANIFOLDS OF EMOTIONAL SPEECH

Jangwon Kim, Sungbok Lee and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab (SAIL)
University of Southern California, Los Angeles, California, USA
jangwon@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

ABSTRACT

This study explores manifold representations of emotionally modulated speech. The manifolds are derived in the articulatory space and two acoustic spaces (MFB and MFCC) using isometric feature mapping (Isomap) with data from an emotional speech corpus. Their effectiveness in representing emotional speech is tested based on the emotion classification accuracy. Results show that the effective manifold dimensions of the articulatory and MFB spaces are both about 5 while being greater in MFCC space. Also, the accuracies in the articulatory and MFB manifolds are close to those in the original spaces, but this is not the case for the MFCC. It is speculated that the manifold in the MFCC space is less structured, or more distorted, than others.

Index Terms— Emotional speech, Isomap, manifold, acoustic feature, articulatory feature

1. INTRODUCTION

Speech events have been analyzed and modeled using various sets of signal feature descriptions. For example, articulatory measurement features have shown how the articulatory movement dynamics are shaped during emotional expression [1]. Similarly, acoustic information pertaining to emotional speech has been described by spectral features, for instance using Mel-frequency cepstral coefficients (MFCCs) and Mel-frequency filter banks (MFBs). All these articulatory and acoustic signal (measurement) descriptions are on feature spaces that vary in their size, as well as redundancy in description. It has been claimed that the speech space, both articulatory and acoustic, can be represented in a lower dimensional feature space than what is obtained with conventional feature descriptions [2] [3]. This claim is based on the innate constraints of human speech production, rules that underlie the creation of linguistically meaningful sounds, and the physiological limits in the vocal tract. Since the modulation of speech source signal results from the combination of articulator motions, these constraints may limit the variation patterns in the speech signal. In this sense, speech data samples in the original signal feature space can indeed be represented in a lower dimensional space that reflects the actual underlying structure. A previous study presented experimental evidence that, for Linear Predictive Coding (LPC) coefficients and MFBs, the underlying structures were best approximated by three to five dimensions [2].

Recently, manifold methods for representing informative signatures in the speech signal have been studied. A manifold is a mathematical space created by the information of distance among a certain amount of data samples. It allows considerable data reduction while retaining critical information of the modulation patterns. Isometric feature mapping (Isomap) [4] is one of the most popularly used algorithms to discover a manifold. A previous study showed that the Isomap-learning based manifold spaces were capable of retaining the distinctive information between phone classes [5].

This paper expands the scope of these previous studies to emotional speech. The main goal is to examine the effectiveness of manifolds for the representation of emotionally modulated speech, both in the articulatory and acoustic domains. In previous studies [1] [6], more peripheral movements of articulators for expressive speech production were discovered. Also, four basic emotions, such as angry, neutral, sad and happy, have distinctive probability distribution functions (PDF) in the spectral feature space [7]. If a manifold is capable of retaining the emotional information of the original feature space, the PDFs between emotions will still be distinctive from each other in the manifold space.

Our experimental investigation considers the following acoustic and articulatory data: MFCCs, MFBs and the trajectory measurements of articulators, such as tongue tip, lower lip and jaw. Manifold representations derived using Isomap were evaluated using emotion classification rate and confusion and compared to those obtained in the original feature spaces.

This paper is organized as follows. First, Isomap is described briefly in section 2. Next, the databases used and the original feature extraction methods, as well as the training and testing methods of four basic emotion models, are explained in section 3. Experimental results and discussions are provided in section 4. Finally, the conclusion of this study and future work are discussed in section 5.

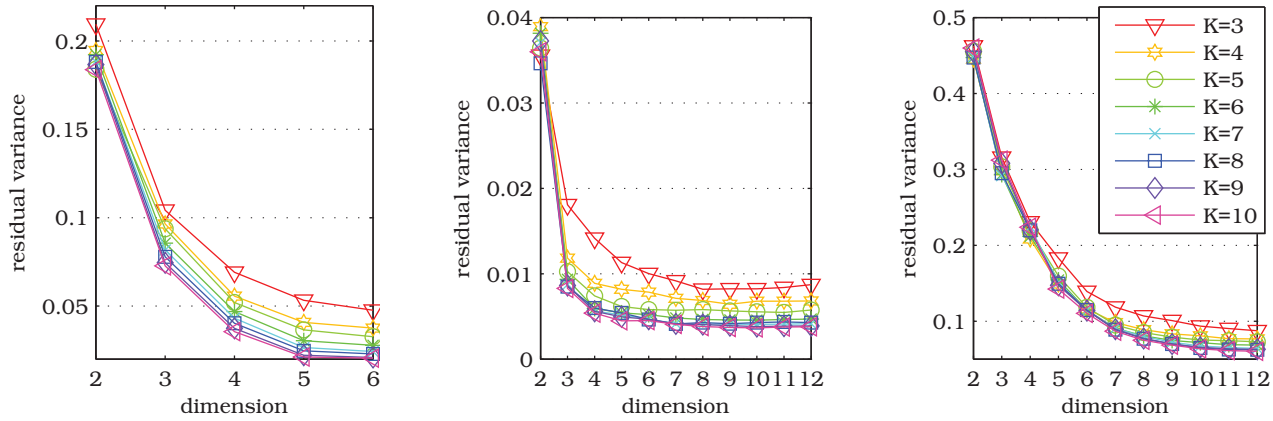
2. ISOMETRIC FEATURE MAPPING

Isomap is an unsupervised learning algorithm which uses geodesic distance on multidimensional scaling (MDS) [4]. This algorithm provides residual variance, which is calculated as below.

$$E_d = 1 - R^2(D_G, D_Y) \tag{1}$$

where E_d is residual variance; D_G is an approximated geodesic distance matrix which is estimated by computing shortest path distances on a manifold; D_Y is a Euclidean distance matrix in the embedded feature space; R is the standard linear correlation coefficient, taken over all entries of D_G and D_Y . E_d was used as an indicator of how well original features fit Isomap better, because a lower residual variance indicates less metric distortion.

The success of finding a manifold depends on how a suitable number of neighborhoods, or K, is chosen. In theory, Isomap guarantees that the intrinsic geometric structure of a manifold is recovered better asymptotically, as K increases [4]. However, a large K causes heavy computational cost. Therefore, an optimized K value, which is cost effective, is generally used. E_d was used for estimating the proper K.



(a) Articulator features (b) MFB (c) MFCC

Fig. 1. Residual variance in low dimensional space from three different original feature spaces, such as articulatory feature space, MFB space and MFCC space. Note the tendency of residual variance of MFCC to saturate in the higher dimensions than articulatory features and MFB. This implies the manifold space from MFCC is less structured than others.

3. METHODS

3.1. Dataset

The USC Electromagnetic Articulography (EMA) emotion database was used for this study [1]. It has both speech waveforms at a 16 kHz sampling rate and synchronized articulatory movement measurements on the x axis and the y axis along the mid-sagittal view of tongue tip, lower lip and jaw at a 200 Hz sampling rate. Articulatory position values were aligned to speech automatically by forced-alignment using hidden Markov models of English monophones. A total of 680 utterances were spoken by three native speakers of American English: two females and a male. One of the females, JN had vocal training in theatre/acting. 10 sentences were commonly spoken by all talkers, and each sentence was repeated 5 times for each of the four basic emotions: neutral, angry, sad and happy.

Only a subset of the corpus (120 utterances of JN) was used in this initial study (3 utterances $\times$ 10 sentences $\times$ 4 emotions). One reason for this was the innate computational complexity of Isomap. Dijkstras algorithm, with Fibonacci heaps with K neighborhoods, which was used for the calculation of the shortest-paths distance matrix is O(kN^2logN), and the MDS eigenvalue calculation is O(N^3) [8]. These two processes require high computation time.

Also, it is necessary to use the frames which are more likely colored by emotion for this study. A previous study showed both that the second half of a sentence is more consistently influenced by emotion and that verbs provide less emotional information than adjectives or nouns [9]. So, the first words at the beginning of utterances, which are all pronouns, and verbs were excluded in this study. Also, we assumed that prepositions and articles were too short to include emotional information in a consistent manner, so they were excluded as well. Finally, we chose total 22 words: GRANDMOTHER, PHONE, DOCTOR, SCAR, FOAM, ANTISEPTIC, FATHER, DEAF, TANTAMOUNT, ISOLATION, ECHO, VOICES, SOUND, SHOES, COMPANY, FUTURE, PICTURE, DRESS, ASIA, LONG, VERY and SAME and then used the frames within boundaries of those words.

3.2. Feature extraction for original feature space

For the articulatory domain, position values on the x axis and the y axis of tongue tip, lower lip and jaw (6 features) are directly extracted from the EMA database. Since syllable level is considered as the minimum duration for conveying emotion, we applied a 100 millisecond Hamming window, which is close to the shortest average syllable duration of words, with 50 millisecond shifting. The mean of samples in a window was used as the articulatory feature.

For the acoustic domain, two sets of speech spectral features are used from the speech signal down sampled to 8 kHz. One set is 12th-order MFCCs, and the other is 12th-order MFBs. They were extracted using the HTK toolkit [10]. DC mean normalization and first order preemphasis with a coefficient of 0.98 were applied as preprocessing. For fair comparison, we used identical window settings with articulatory feature extraction.

3.3. Training and Testing of Emotion Models

80 utterances, two from each sentence of each emotion, were used for training each emotion model, and the other utterances were used for testing. Each emotion class was modeled by Gaussian Mixture Models (GMMs). For training emotion models, we used NETLAB, which provides MATLAB functions for parameter initialization by k-means algorithm, and expectation-maximization algorithms [11]. Each frame was classified based on log-likelihood ratio. K-means algorithm, which was used for determination of initial parameters of GMMs, causes some variation of them. However, the standard deviations of emotion classification results repeated 10 times in the original feature space were small enough to be ignored: less than 0.02 in the articulatory feature domain and less than 0.01 in both MFCC and MFB domains. The number of mixtures was chosen experimentally at the performance elbow point: 8 mixtures for articulation-feature-based models, 20 mixtures for MFB-based models and 8 mixtures for MFCC-based-models.

Table 1 shows the confusion matrix of each feature type at their elbow points. These results will be compared with those of Isomap embedded space.

Table 1. Confusion matrix in the original feature space of each feature type. Bold characters indicate highest classification rate. (Ang = Angry, Neu = Neutral, Sad = Sadness, Hap = Happy, T = Total number of samples, CR = Classification rate)

			Classified					
			Ang	Neu	Sad	Hap	T	CR
Ground truth	ARTI	Ang	**183**	25	66	30	304	.60
		Neu	5	**259**	0	7	271	.96
		Sad	7	96	**195**	3	301	.65
		Hap	80	26	4	**203**	313	.65
	MFB	Ang	**121**	43	57	83	304	.40
		Neu	15	**139**	109	8	271	.51
		Sad	8	91	**193**	9	301	.64
		Hap	133	26	24	**129**	312	.41
	MFCC	Ang	**121**	47	66	70	304	.40
		Neu	5	**173**	74	19	271	.64
		Sad	8	97	**188**	8	301	.62
		Hap	91	27	23	**171**	312	.55

4. RESULTS AND DISCUSSIONS

4.1. Residual variances

Residual variance allows us to find the inflection point where a suitable neighborhood size and dimensionality of the low dimensional space are decided in general. Figure 1 shows residual variances for different settings of Ks and dimensions. A suitable K is determined to be between 3 to 9 in practice [12]. In this figure, 7 to 8 neighborhoods appear to be adequate enough to convey geodesic distance information, so 8 neighborhoods are used in this study.

Figure 1 also indicates that elbow points exist at 3 to 5 dimensions and their residual variances are not much different from those of higher dimensions when suitable Ks are used. These results indicate that 3 to 5 dimensional spaces by Isomap preserves geodesic information as much as higher dimensional spaces.

4.2. Emotion information in low-dimension feature space

In this section, the capability of representing emotional information in low dimensional feature space was examined. It was done by comparing classification performances in the low dimensional feature space with those in the original feature space. The number of mixtures in the low dimensional feature space was identically set with that in the original feature space for fair comparison.

Figure 2 shows emotion classification results in the low dimensional feature spaces. In this figure, classification rates do not increase significantly after 4 dimensions for both articulatory features and spectral features. Even when we tested with different numbers of mixtures ranging from 2 to 32, the elbow points appeared consistently in 3 to 5 dimensions, mostly in 4 dimensions. It may indicate that about 4 dimensions are optimum to retain distinctive PDF information between basic emotions in the manifold space. In general, the motions of jaw and lower lip are highly correlated, which may influence the saturation in 4 dimensions for articulatory features. Interestingly, a previous study reported that the inflection point of phone group classification results in a manifold also appeared in 3 to 5 dimensions [5].

For MFBs and articulatory features, manifold spaces retain emotional information almost as much as the original feature

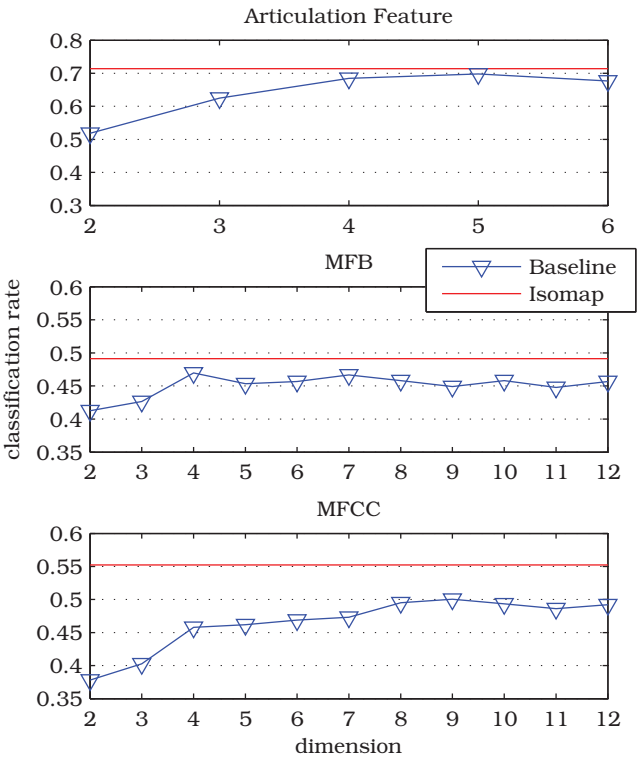


Fig. 2. Emotion classification results of three feature types in manifold. Baseline indicates emotion classification rate in original feature space.

space: 2.2% lower for MFBs and 3% lower for articulatory features. However, the manifold space from MFCCs retains relatively much smaller emotional information: 9.4% lower. These results may indicate that the manifold space from MFCCs is less structured than those from MFBs and articulatory features. On the other hand, the classification rates of embedded spaces of the identical dimension sizes as their original feature space are inferior to those in the original feature spaces. It may be due to the fact that the embedding process has an information loss problem from the distortion of geodesic distance information as shown by the residual variances in Figure 1.

Table 2 shows the confusion matrix for the four basic emotion classes in the 4-dimensional feature space. As we can see in Table 1 and Table 2, confusion patterns in the classification in the original feature spaces and Isomap feature spaces are similar overall. This supports the observation that the manifold space retains distinctive information between basic emotions well.

5. CONCLUSION AND FUTURE WORK

In this paper, we explored the effectiveness of using manifold representations for analyzing emotional speech utterances. The representation capability was evaluated using emotion classification rates and the confusion matrix. Results show that 3 to 5 feature dimensions are the optimum numbers to convey emotional information in manifold spaces derived from three different feature sets derived from

Table 2. Confusion matrix in 4 dimensional space of each feature set. Bold characters indicate highest classification rate. (Ang = Angry, Neu = Neutral, Sad = Sadness, Hap = Happiness, T = Total number of samples, CR = Classification rate)

			Classified					
			Ang	Neu	Sad	Hap	T	CR
Ground truth	ARTI	Ang	**169**	31	86	18	304	.56
		Neu	7	**252**	4	8	271	.93
		Sad	6	71	**222**	2	301	.74
		Hap	75	66	10	**161**	312	.52
	MFB	Ang	**108**	69	69	58	304	.36
		Neu	10	**125**	123	13	271	.46
		Sad	2	86	**202**	11	301	.67
		Hap	105	46	39	**122**	312	.39
	MFCC	Ang	**93**	72	88	51	304	.31
		Neu	18	**142**	97	14	271	.52
		Sad	3	103	**181**	14	301	.60
		Hap	106	49	32	**125**	312	.41

speech articulatory and acoustic data. Also, the classification results showed that the manifold spaces derived from MFBs and articulatory features were close to those obtained from the original feature spaces. The manifold space from MFCCs shows relatively higher information loss, which indicates that it has a more noisy underlying structure than in the original feature space.

While the results from this initial study point to the potential effectiveness of a lower dimensional manifold representation for emotional speech, more detailed experiments are required to further validate the observations made here. First, we need more and further detailed articulatory data; we hope the creation of a more extensive corpus of emotional speech with EMA and Magnetic Resonance Imaging can help in this regard for future studies. Second, for the manifold based embedding algorithms, it is difficult to interpret what physical constraint corresponds to each parameter, so this remains an interesting open question. Finally, it is necessary to develop more sophisticated manifold learning algorithms that are capable of retaining more information with less distortion. Our future work will include these directions.

6. REFERENCES

[1] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *Proceedings of Eurospeech*, 2005, pp. 497–500.

[2] R. Togneri, M.D. Alder, and Y. Attikiouzel, "Dimension and structure of the speech space," *Communications, Speech and Vision, IEE Proceedings I*, vol. 139, no. 2, pp. 123–127, April 1992.

[3] A. Jansen and P. Niyogi, "Intrinsic fourier analysis on the manifold of speech sounds," in *ICASSP*, 2006, vol. 1, pp. 241–244.

[4] J.B. Tenenbaum, V.D. Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, December 2000.

[5] A. Errity, J. McKenna, and B. Kirkpatrick, "Manifold learning-based feature transformation for phone classification," in *Proc. of the ISCA Tutorial and Research Workshop on Nonlinear Speech Processing (NOLISP)*, Paris, France, May 2007, pp. 43–46.

[6] D. Erickson, C. Menezes, and A. Fujino, "Some articulatory measurement of real sadness," in *Proceedings in Interspeech*, October 2004, pp. 1825–1828.

[7] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Proceedings of Interspeech ICSLP*, 2004, pp. 889–892.

[8] V.D. Silva and J.B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," *Advances in Neural Information Processing Systems (NIPS) 15*, vol. 15, pp. 705–712, 2003.

[9] M. Bulut, S. Lee, and S. Narayanan, "Analysis of emotional speech prosody in terms of part of speech tags," in *Proceedings of Interspeech ICSLP*, 2007, pp. 626–629.

[10] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory, December 2006.

[11] C. Bishop and I. Nabney, *Netlab neural network software*, http://www.ncrg.aston.ac.uk/netlab/index.php.

[12] S. Weng, C. Zhang, , Z. Lin, and X. Zhang, "Mining the structural knowledge of high-dimensional medical data using isomap," *Medical and Biological Engineering and Computing*, vol. 43, pp. 410–412, 2005.