INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Generating Gestural Scores from Acoustics through a Sparse Anchor-Based Representation of Speech

Christopher Liberatore, Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, United States
cliberatore@cs.tamu.edu, rgutier@cs.tamu.edu

Abstract

We present a procedure for generating gestural scores from speech acoustics. The procedure is based on our recent SABR (sparse, anchor-based representation) algorithm, which models the speech signal as a linear combination of acoustic anchors. We present modifications to SABR that encourage temporal smoothness by restricting the number of anchors that can be active over an analysis window. We propose that peaks in the SABR weights can be interpreted as "keyframes" that determine when vocal tract articulations occur. We validate the approach in two ways. First, we compare SABR keyframes to maxima in the velocity of electromagnetic articulography (EMA) pellets from an articulatory corpus. Second, we use keyframes and SABR weights to build a gestural score for the VocalTractLab (VTL) model, and compare synthetic EMA trajectories generated by VTL against those in the articulatory corpus. We find that SABR keyframes occur within 15-20 ms (on average) of EMA maxima, suggesting that SABR keyframes can be used to identify articulatory phenomena. However, comparison of synthetic and real EMA pellets show moderate correlation on tongue pellets but weak correlation on lip pellets, a result that may be due to differences between the VTL speaker model and the source speaker in our corpus.

Index Terms: gestural scores, articulatory inversion, vocal tract model, sparse coding

1. Introduction

Physical models use detailed two or three dimensional meshes to simulate the configuration of the vocal tract and its articulatory parameters. These models can generate synthetic acoustics for a specific articulatory configuration by measuring the cross-sectional area of the model and then computing the resulting filter resonances. To produce continuous speech, these models require a motor control program that emulates motions and phenomena seen in the vocal tract, such as coarticulation. An example of such model is VocalTractLab (VTL) [1, 2]. VTL presents the vocal tract as a set of seven distinct 3-dimensional meshes and uses a control program inspired by the Task Dynamics framework [3] to model vocal tract dynamics. Task Dynamics models vocal tract motions as a collection of "gestures," or a period when an articulator is moving towards an articulatory target. Generally, a gestural control program (known as a "gestural score") is built by a human expert and requires *a priori* knowledge on the desired phonemic content. As noted by Nam et al. [4], however, estimating gesture timings from speech acoustics is difficult.

This paper examines the possibility of deriving gestural scores directly from acoustics using a recently proposed sparse, anchor-based representation (SABR) of speech [5]. SABR decomposes speech as a weighted sum of speaker-specific acoustic anchors. By choosing phoneme centroids as anchors, and by ensuring that the decomposition is sparse, the weight matrix describes *what* phonemes are uttered, and *when*. If SABR anchors can be mapped into articulatory configurations, can the SABR weighs be used to estimate gestural scores?

To answer this question, we present an iterative method to promote temporal smoothness on the SABR weights by imposing constraints on how many anchors can be active over an analysis window. We then extract features from the SABR weight matrix that indicate when articulatory motions begin—a concept we call "keyframes". From the SABR weights and corresponding keyframes, we generate gestural scores for VTL. We show that SABR weights can predict the onset and offset timings of gestures by comparing extracted keyframes with EMA data from an articulatory corpus. Finally, we use keyframes and SABR weights to generate gestural scores for VTL, and compare synthetic EMA trajectories generated by VTL against EMA trajectories from the source speaker.

The rest of the paper is organized as follows. Section 2 reviews previous work on gestural scores generation and sparse representations of speech; it also provides a brief overview of the VocalTractLab physical model. Section 3 describes the original SABR model and proposes three modifications that allow it to generate gestural scores: windowed generation of SABR weights, grouping weights by manner of articulation, and using multiple anchors to represent phonemes. This section also describes how we extract keyframes from SABR weights and generate gestural scores for VTL. Finally, we review the performance of the keyframe selection algorithm as well as the gestural score generation, and discuss results and future work.

2. Background

The Task Dynamics model of Saltzman and Munhall [3] describes vocal tract motion as a constellation of "gestures", or periods of time articulators of the vocal tract are moving towards a place and degree of constriction. Coordinating these gestures allows a speaker to reach configurations necessary to produce phonemes. The Task Dynamics model explains many observed phenomena of vocal tract motion, such as coarticulation and speaking rate variations [6]. Under the Task Dynamics model, gestures can also provide a means to perform a form of articulatory inversion; it has been shown that gestures can be applied to various speech applications, from speech recognition [7] to speech synthesis [8].

A few methods have been proposed to automatically transcribe gestural scores from a source utterance. Nam et al.

http://dx.doi.org/10.21437/Interspeech.2016-1336

[4] used a phomeic transcription from a source utterance to drive the TADA vocal tract model (based on Task Dynamics, [9]). The authors then used an iterative time warping to align the onset and offset times of the gestural score by comparing using a warping function between the synthetic TADA acoustics and source utterance. They found that their proposed iterative procedure produced lower log-spectral distances than that of standard dynamic time warping. In related work, Steiner and Richmond [10] proposed a gestural score generation for the VTL model using a source phonemic transcription. The authors posed the score generation problem as that as determining the best path through a transition network modeled after a finite state automata. They used Viterbi search to determine the path that maximized the correlation between synthetic EMA trajectories generated by VTL and those of a source speaker. Their results were preliminary, but they found moderate correlation with the synthetic and source EMA trajectories.

A related method to SABR is the Temporal Decomposition (TD) of Atal et al. [11]. TD represents an utterance Y as a linear combination of acoustic basis functions Φ: $Y = A\Phi$. Unlike SABR weights, the basis functions Φ in TD capture a set of acoustic parameters in addition to the time frame those parameters are observed; the amplitudes, A, modify the magnitude of those parameters independent of time. The temporal component of the basis functions allowed the authors to estimate when "speech events" occurred.

2.1. VocalTractLab

VocalTractLab (VTL) [1] is a physical model of the vocal tract that includes both a synthesizer and a gesture-based control model [2]. The model represents the vocal tract using seven meshes and controls articulations of these surfaces with 21 "tract parameters". VTL's gestures are represented as specific locations of each of these control parameters; degrees of constriction are not a direct parameter in the VTL gestures. Given a gestural score, VTL models motion between successive gestures as a 10th-order dampened system. For more details on VTL's control model, please see [12].

3. Methods

The proposed methods are based on our prior work on the development of SABR (Sparse Anchor-Based Representation) [5]. SABR is a speech decomposition method that uses a set of speaker-dependent acoustic vectors as "anchors." Given a source utterance $\boldsymbol{X} \in \mathbb{R}^{f \times n}$ and a set of speaker anchors $\boldsymbol{A} \in \mathbb{R}^{f \times a}$, SABR represents an utterance as:

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{W} \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{a \times n}$ is a set of weights, one channel per acoustic anchor, of length equal to that of the source utterance. SABR uses a sparse, nonnegative least squares (NNLS) solver (Lasso [13]) to solve for $\boldsymbol{W}$, resulted in a *speaker-independent* representation of the utterance.

3.1. Improvements to SABR

A limitation of the original SABR algorithm is that it considers each acoustic frame independently of all others. As a result, phonemes with more turbulence tend to have less temporal stability in SABR weights. The following sections describe modifications to the SABR algorithm and anchor set to improve temporal stability of the representation, which is essential for gestural score generation.

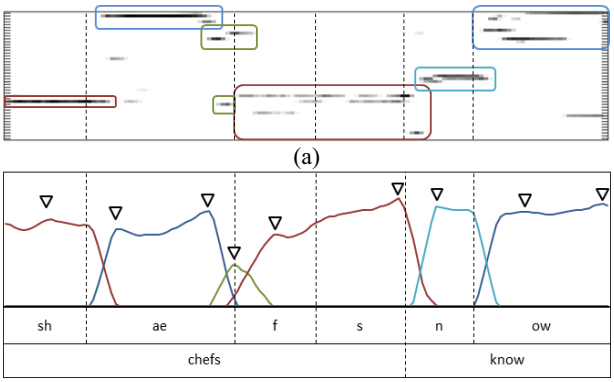


Figure 1: SABR modifications and keyframe detection. (a) Iterative SABR weights. (b) The same weights as the previous figure, but grouped by manner of articulation. The ground-truth phonemic and orthographic transcriptions are included for reference. Triangles identify peaks of the manner channels, which we call "keyframes".

3.1.1. *Windowed SABR*

Typically, the less sonorant a phoneme is, the lower the temporal stability of its SABR weights (see figure 1a); this lack of stability hampers interpreting SABR in an articulatory context. To address this, we modified SABR to enforce sparsity over an analysis window spanning multiple acoustic frames. Specifically, given a window d centered on frame i, we add up the SABR weights for each anchor and select the two anchors a_1 and a_2 with the highest cumulative weights:

$$\underset{a_1,a_2}{\operatorname{argmax}} \sum_{n=i-d}^{i+d} \left(\boldsymbol{W}_{a_1,n} + \boldsymbol{W}_{a_2,n}\right) \tag{2}$$

We solve again for the weights at frame i using NNLS but only consider the anchors chosen in equation (2):

$$\min \left|\left|\boldsymbol{X}_i - [\boldsymbol{A}_{a_1}\, \boldsymbol{A}_{a_2}][\boldsymbol{W}_{a_1,i}\, \boldsymbol{W}_{a_2,i}]\right|\right| \tag{3}$$

This step is intuitively similar to Modified Restricted Temporal Decomposition [14], in that it limits the number of active anchors at any one frame and takes into account prior and future weights at a given frame.

3.1.2. *Manner filtering*

Even after the temporal smoothing step in equations (2) and (3), anchors with similar manner of articulation may still be confounded. Consequently, we group weight channels according to the manner of articulation[1] of their respective anchor; each "manner channel" is the sum of the SABR weights that have the same manner of articulation. For a set of manners Γ, the manner weights $\boldsymbol{M} \in \mathbb{R}^{|\Gamma| \times n}$ are:

$$\boldsymbol{M}_{\gamma,i} = \sum_{\gamma \in \Gamma} \boldsymbol{W}_{m,i}\ s.t.\ m = \{\forall a \in A : man(a) = \gamma\} \tag{4}$$

where $man(a)$ returns the manner of articulation of the anchor a. The resulting manner channels have more temporal stability than windowed SABR weights; see Figure 1(a-b). We take advantage of this stability to update the SABR weights as follows. Each channel of $\boldsymbol{M}$ is sparse, comprised of

[1] We classified each anchor according to one of the following manners: vowel, stop, fricative, approximant, or occlusive.

Table 1: rules for generating multimodal SABR anchors

Phoneme class	States	Rationale
Voiced stops	2	Stop and release segments
Unvoiced stops	3	Stop, release, and optional aspiration segment
Diphthongs	0	Removed, as these transitions between two vowel states
Affricatives	0	Removed these, as they transition between stop and fricative states

contiguous, nonzero regions. For each contiguous region R, we examine the SABR weight channels in that region and use the manner weights to replace the SABR weights as follows. For all SABR channels i:

$$\boldsymbol{W}_{i,R} = \begin{cases} \boldsymbol{M}_{man(i),R} & \text{if } i = argmax_i \, \Sigma \boldsymbol{W}_{i,R} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

If a SABR channel has the highest cumulative weights in the region R , we assign the SABR channel in that same region the weights from the manner weights $\boldsymbol{M}$. Otherwise, we set the weights for that anchor and spanning that region to zero.

3.1.3. Multimodal SABR models

The original SABR implementation uses phoneme centroids as anchors (i.e., one anchor per phoneme). When used to generate gestural scores, this model is limited because not all phonemes can be represented by a single segmental state. As an example, consider stops, which have closure, stop, and release segments—these segments all represent different acoustic segments of the same stop gesture.

To address this issue, we modified our anchor-building procedure in two ways. First, we made rules for either removing or expanding our anchor set based upon segmental properties of the phonemes the anchors represent—see Table 1. Phonemes with multiple segments received an increase in the number of anchors accordingly. However, if the segments of that phoneme exist outside of that phonemic context (consider the affricative /dʒ/, which transitions between /d/ and /ʒ/), we remove that anchor entirely.

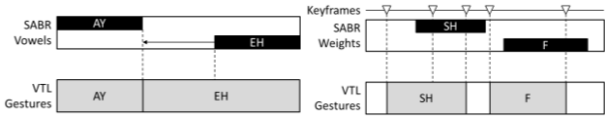
Second, when a phoneme was represented by n anchors, we used Ward's Method [15] as a criterion to cluster the anchors within that phoneme. Our clustering only had two levels—a root level, and a leaf level with n leaves that represented our cluster centers. Once we computed the cluster centers we use the closest sample in the dataset as the anchor.

3.2. Selecting keyframes using SABR weights

We use the weights that result from the modified SABR procedure to infer the timing of articulatory transitions. Namely, we assume that a local maximum in a SABR weight indicates a transition of the vocal tract towards a new gestural target. In deference to the animation literature, we refer to these SABR peaks as "keyframes." These keyframes are the basis for selecting gesture timings in our approach. Specifically, we choose keyframes as the union of all peaks of each SABR weight channel. Given a set of a anchors, we define the SABR keyframes as:

$$K_{SABR} = \bigcup_{i=0}^{a} peaks(\boldsymbol{W}_i) \quad (6)$$

where W_i is the ith frame of W and $peaks$ returns the set of locations of positive-valued positions where the second derivative of the channel is negative. We found that first filtering the weights with a smoothing filter of width 15ms made the peaks interpretable.



(c) (d)

Figure 2: Gestural score transform. (c) Using SABR weights to build the VTL vowel gestures. (d) Using SABR weights and extracted keyframes to build a non-vowel VTL gesture tier.

3.3. Gestural score transform

To generate a gestural score, we define a phoneme map $P: A \rightarrow G, T$, where A is a SABR anchor, G is a VTL gesture, and T is a VTL "tier". We built the map P manually, according to known places and manners of articulation for each anchor (based on specifications in TADA [9]). VTL defines two classes of gestures: vowel gestures and consonant gestures. Vowels belong to one tier, whereas consonants are associated to three different tiers and define lip, tongue tip, and tongue body gestures[1].

For *vowels*, we assign gestures at each frame by finding the maximum SABR vowel weight at that frame. If there are no vowels with weights at that frame, we look ahead until the first frame with a vowel weight, and propagate that back through the empty frames; see Figure 2(c). This is required since VTL vowel gestures are always active [12].

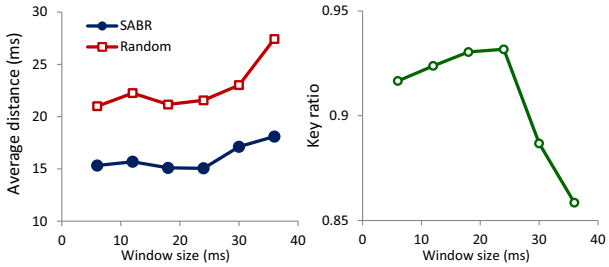
For each *consonant tier* T, we consider the SABR weight channels that belong to tier. We use the phoneme map P to map each weight channel to a corresponding gesture and tier, as each weight channel is associated with a specific anchor. Since weight channels are sparse, we consider each group of contiguous nonzero weights an expression of a gesture. We then find two keyframes within the time range of that gesture to map the onset and offset times of the gesture G associated with that weight channel. The onset keyframe is the first keyframe before the first positive weight value, and the offset of the gesture is the last keyframe within that group of weights; see Figure 2(d).

4. Data

We validated the method on an articulatory corpus from a North American English speaker, described elsewhere [16]. Seven EMA pellets (upper and lower lips, upper incisor, jaw, tongue tip, tongue middle, tongue rear) were placed in the subject's vocal tract, and samples were collected at a rate of 200 Hz. The upper incisor was used as a reference point for the other 6 points and the X and Y positions for these points were tracked, resulting in 12 EMA channels. The corpus included 344 sentences from the Glasgow Herald corpus. Of these utterances, we selected a subset of 20, which maximized phoneme balance, to conduct our experiments.

As in our prior work [5], we used STRAIGHT [17] to extract spectral envelopes from the source speaker, then computed 24 Mel-Frequency Cepstral Coefficients for use in our acoustic SABR model.

[1] VTL has four additional tiers relating to the degree of velum opening, glottal shape, pitch, and lung pressure. We do not consider them here as we are more concerned with supraglottal motions.



(a) (b)

Figure 3: keyframe distance and ratio comparisons. (a) the mean keyframe distance. SABR keys have significantly lower average distances than randomly-chosen keys. (b) the ratio of the number of SABR keyframes to the number of EMA keyframes.

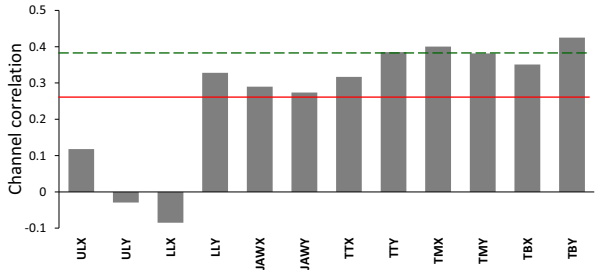


Figure 4: correlations between synthetic and collected EMA data. All channels had an average correlation of $p = 0.26$, but the tongue channels (highlighted) had an average of $p = 0.38$.

5. Results

5.1. Keyframe interpretation

As a first experiment, we compared our estimated keyframes K_{SABR} against those extracted from EMA pellets. For a given pellet p, its EMA trajectory data, E_p, has both X and Y channels. We compiled EMA keyframes, K_{EMA} as:

$$K_{EMA} = \bigcup_{\forall p} peaks(\Delta E_p) \tag{7}$$

where ΔE^p is the Euclidean distance between successive frames of E^p.

We hypothesized that if SABR keyframes capture articulatory changes there should be the same number of SABR keyframes as there are EMA keyframes:

$$|K_{SABR}| \approx |K_{EMA}| \tag{8}$$

Additionally, we hypothesized that the positions of SABR and EMA keyframes should be approximately the same, that is:

$$\forall k \in K_{SABR}, \min(||k - K_{EMA}||) \to 0 \tag{9}$$

We extracted SABR keyframes over a range of window sizes and evaluated eqs. (8) and (9). For a set of SABR keyframes extracted for a window size, we evaluated eq. (8) by computing the ratio of the cardinality of SABR and EMA keyframe sets. For eq. (9), we computed the average distance from each SABR keyframe to the nearest EMA keyframe. As a measure of baseline performance, we also compiled a random set of frames, K_{RND}, of the same cardinality as K_{SABR}, and evaluated their performance in the same manner as the SABR keyframes.

Results are shown in Figure 3. A window size of 24 ms minimizes the average distance to the nearest keyframe to 15.06 ms. The randomly-selected keyframe had significantly higher distances to EMA keyframes ($p < 0.01$) than the SABR keyframes, suggesting that the SABR keyframes are not finding articulatory changes by chance. Interestingly, the ratio of SABR to EMA keyframes also peaks for a window size of 24 ms; see Figure 3 (b).

5.2. Gestural score estimation

In a second experiment, we used the gestural score generation algorithm from section 3.3 to build scores from the 20 source utterances. For a given gestural score, VTL is able to synthesize EMA pellet trajectories by tracking vertices of the underlying anatomy meshes. To measure how well our generation algorithm worked, we computed the correlation between synthetic and source EMA data.

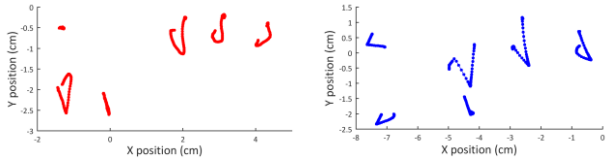
Using the window size that maximized the key ratios and minimized the average keyframe distance, we use the algorithm in section 3.3 to generate gestural scores from VTL. To account for the fact that gestures are active before motion begins, we offset the gesture activation times by 75 ms, similar to what Birkholz reported in [18]. Using a window size of 24 ms, the synthetic gestural scores had an average correlation across all channels of $\rho = 0.26$. However, the lip EMA pellets were for the most part uncorrelated; when we examined only the correlation of the tongue pellets, we found a higher average correlation of $\rho = 0.38$—see Figure 4.

6. Discussion

In this paper, we proposed a technique to generate gestural scores for VocalTractLab from "keyframes" extracted from SABR weights. This technique relied on some modifications to the SABR method to make the representation more temporally stable. In a first experiment, we found that the SABR keyframes were, on average, within 15 ms of the observed articulatory events. In a second experiment, we measured the correlation of the synthetic EMA trajectories from the VTL model driven with our gestural score method with trajectories from a source speaker. We found that there was moderate correlation on tongue channels, but less so on lip and jaw channels.

Several factors account for the correlation performance. First, we did not adjust the articulatory effort for each gesture; the VTL model reached gestural targets more quickly than the source speaker and transitions were not as smooth (Figure 5). Secondly, we used the default VTL model, and EMA pellet positions may not be optimal. Third, adjacent phonemes with the same manner of articulation were combined into one gesture because of our manner filtering step.

Future work includes improving the SABR weight smoothing procedure by incorporating smoothing constraints in the NNLS step. As keyframes provide some knowledge of articulatory events, it may be beneficial to use these features in a small-model speech synthesis context, such as SABR voice conversion.



(a) (b)

Figure 5: EMA trajectories for the word "chefs" from source speaker (a) and VTL (b).

7. References

[1] P. Birkholz. (2010). *Vocal Tract Lab*. Available: www.vocaltractlab.de

[2] P. Birkholz*, et al.*, "Construction and control of a three-dimensional vocal tract model," in *ICASSP*, 2006, pp. 873-876.

[3] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological psychology,* vol. 1, pp. 333-382, 1989.

[4] H. Nam*, et al.*, "A procedure for estimating gestural scores from speech acoustics," *Journal of the Acoustical Society of America,* vol. 132, pp. 3980-3989, 2012.

[5] C. Liberatore*, et al.*, "SABR: Sparse, Anchor-Based Representation of the Speech Signal," in *Interspeech*, Dresden, Germany, 2015, pp. 4250-4254.

[6] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica,* vol. 49, pp. 155-180, 1992.

[7] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *Journal of the Acoustical Society of America,* vol. 111, pp. 1086-1101, 2002.

[8] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *Journal of the Acoustical Society of America,* vol. 137, pp. 433-446, 2015.

[9] H. Nam*, et al.*, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *Journal of the Acoustical Society of America,* vol. 115, p. 2430, 2004.

[10] I. Steiner and K. Richmond, "Towards unsupervised articulatory resynthesis of German utterances using EMA data," in *Interspeech*, Brighton, United Kingdom, 2009, pp. 2055-2058.

[11] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," in *ICASSP*, 1983, pp. 81-84.

[12] P. Birkholz*, et al.*, "Control concepts for articulatory speech synthesis," in *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2008.

[13] J. Mairal*, et al.*, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research,* vol. 11, pp. 19-60, 2010.

[14] P. C. Nguyen*, et al.*, "Modified restricted temporal decomposition and its application to low rate speech coding," *IEICE TRANSACTIONS on Information and Systems,* vol. 86, pp. 397-405, 2003.

[15] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association,* vol. 58, pp. 236-244, 1963.

[16] D. Felps*, et al.*, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 20, pp. 2301-2312, 2012.

[17] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology,* vol. 27, pp. 349-353, 2006.

[18] P. Birkholz*, et al.*, "Model-based reproduction of articulatory trajectories for consonant–vowel sequences," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1422-1433, 2011.