Can unquantised articulatory feature continuums be modelled?

Odette Scharenborg[1] *and Vincent Wan*[2]

[1] Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[2] Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK
O.Scharenborg@let.ru.nl, V.Wan@dcs.shef.ac.uk

Abstract

Articulatory feature (AF) modelling of speech has received a considerable amount of attention in automatic speech recognition research. Although termed ‘articulatory’, previous definitions make certain assumptions that are invalid, for instance, that articulators ‘hop’ from one fixed position to the next. In this paper, we studied two methods, based on support vector classification (SVC) and regression (SVR), in which the *articulation continuum* is modelled without being restricted to using discrete AF value classes. A comparison with a baseline system trained on quantised values of the articulation continuum showed that both SVC and SVR outperform the baseline for two of the three investigated AFs, with improvements up to 5.6% absolute.

Index Terms: articulatory features, speech analysis, speech recognition.

1. Introduction

Articulatory feature (AF) modelling of speech as an alternative to phone modelling has received a considerable amount of attention in automatic speech recognition (ASR) research [1-4]. AF modelling is often considered as the solution to the problem of modelling the variation in speech using the standard ‘beads-on-a-string’ (i.e. using phones) paradigm [5]. AFs are physiologically motivated classes which characterise the essential aspects of articulatory properties of speech sounds for speech perception (e.g., voice, nasality) in a quantised form [1].

Although AFs are termed ‘articulatory’, previous definitions make certain assumptions that are invalid. For instance, in order to train and test an AF-based ASR system it is assumed that articulators ‘hop’ from one fixed position to the next; disregarding the fact that the articulators move continuously from one position to the next. The articulation continuum is quantised and each class is assigned a label (referred to as an AF value) describing the *target* positions of the articulators. ASR systems are then trained and tested on these classes; see e.g. [1-4].

Previous work by [3,4] shows that the least well recognised AFs are those related to tongue position during vowel production and consonantal place of articulation, i.e. AFs of which the possible tongue positions or places of constriction during vowel or consonant production, respectively, seem to lie on an *articulation continuum* from the front to the back or from the top to the bottom of the oral cavity. Consider a vowel: the absolute tongue position for its production is traditionally determined relative to some external reference point [6]. The AF value of that vowel, however, is specified relatively within a group of vowels and not absolutely across all vowels. Therefore, a second vowel with the same absolute position might be assigned a different AF value. This results in a broader distribution of MFCCs associated with the same quantised AF value resulting in an overlap of MFCCs for different AF values. This could possibly explain the somewhat disappointing recognition rates of those AFs.

These issues lead us to the novel idea of modelling the articulation continuum such that one is no longer restricted to using discrete AF values (referred to as modelling of the unquantised articulation continuum); thus eliminating the assumption that articulators ‘hop’ from one fixed position to the next. The question we try to answer is: can articulatory feature continuums be modelled without quantisation? Two methods of modelling the unquantised AF space are compared (section 4). In the first method we train support vector classifiers (SVCs) on the extreme values of an AF continuum and the intermediate places of articulation are inferred. For this to work, it is necessary to establish whether the assumed articulation continuum has an equivalent in the acoustic (MFCC) space: SVCs are binary classifiers but the AFs have multiple quantisation levels so the continuum must be inferable from the extremes. The second method uses support vector regression (SVR) to create a function that describes the articulation continuum. Both are compared to a baseline SVC system trained on quantised values (section 3).

2. Classification systems and material

2.1. Classification systems

Most AF research has been carried out with multilayer perceptrons (MLPs) yielding good performance levels [1-4]. MLPs can assign posterior probabilities between 0 and 1 to points within *uncertainty regions*, i.e. regions where AF value classes overlap, and posterior probabilities of either 1 or 0 to all points outside the uncertainty regions. This implies that if two points in MFCC space are easily classified, e.g. as *high*, they both will be given a posterior probability of 1, making it difficult to indicate whether one point is actually higher than the other.

SVCs [7] make binary decisions by constructing a hyperplane that separates the two classes so that the boundary is geometrically furthest away from both. For each point, SVC assigns a score that lies on a continuum. The score is proportional to the point’s distance from the decision boundary (the point’s classification is usually inferred from the score’s sign). In contrast to posterior probabilities, SVC scores are not restricted to a range of values. Such scores can therefore be used to infer how far a point is along the continuum from one class to another.

Regression is a natural way to model a continuum. SVR [7] is a technique closely related to SVC. It differs from other regression algorithms in a number of ways. Firstly, it is non-parametric so a regression function need not be predefined – the function is estimated by a sum over a set of basis functions that are defined by the kernel. Secondly, most regression algorithms penalise all deviations of the regression function from the data (the goal is to minimise the total penalty). In contrast, SVR does not penalise

[1]The research of Odette Scharenborg was supported by the Netherlands Organization for Scientific Research (NWO) and was carried out while Odette Scharenborg was a visiting scholar at the University of Sheffield.
[2]Vincent Wan was partly supported by the EU 6th FWP IST Integrated Project AMI (Augmented Multparty Interaction, FP6-506811). The authors would like to thank Roger K. Moore for helpful discussions on this work.

until the deviation is greater than a parameter ε. This leads to an "ε-insensitive zone" around the regression function. When ε=0 SVR behaves in much the same way as, for example, MLP regression, by penalising all deviations. Thus SVR has an extra degree of freedom giving it greater potential to create a better model of the continuum by allowing the score for an MFCC to take any value inside the insensitive zone instead of forcing it to a discrete value.

The LIBSVM [8] package was used for both SVC and SVR. In initial experiments, both the polynomial and the RBF kernels were tested on the same task. The RBF kernel showed a better result so the experiments reported here use only that kernel.

2.2. Material

The TIMIT [9] speech corpus was used in this study. It consists of reliably hand labelled and segmented data of quasi-phonetically balanced sentences read by 630 native speakers of American English. TIMIT's standard training and testing division was followed so that sentences and speakers used in training were not used in testing. The training data consists of 3,696 utterances. The test data (excluding the `sa` sentences) consists of 1,344 utterances. AF labels were derived by substituting the frame-level phonemic TIMIT labels with the canonical AF values using a look-up table [4]. The speech was parameterised with 12 MFCC coefficients and log energy, augmented with their first and second derivatives and extended with a context window of ±3 frames resulting in 273-dimensional MFCC vectors.

2.3. The articulatory features

We investigate three AFs that seem to have an articulation continuum: 'fr-back' describes the tongue position on the front-back continuum for vowels using quantised AF values *front, central, back*, and *nil*; 'high-low' describes the tongue position on the high-low continuum for vowels using AF values *high, mid, low*, and *nil*; 'place' describes the place of constriction for consonants using AF values *bilabial, labiodental, dental, alveolar, velar, nil*, and *silence*.

3. The baseline system

A baseline SVC system trained on quantised AF values is used to assess the results of the two methods of modelling the unquantised articulation continuum. Since SVCs can generalise to a small amount of high-dimensional data, not all available training material was used. Instead, a smaller training set was created by randomly selecting 500K frames (i.e. 44.2%) from the full training set while keeping the same prior distribution. Table 1 shows the baseline classification results for each AF separately in terms of accuracy, i.e. the percentage of frames correctly classified. The 'Acc.' column shows the accuracies calculated over all test frames. The accuracies reported in the 'Acc' column are higher than those reported by [3,4]. Since we are primarily interested in how well our new methods are able to model the articulation continuums of the three AFs under investigation, we will compare the performance of the new methods with the baseline system when the *nil* and *silence* frames are discarded from the test set: *nil* and *silence* are not parts of the articulation continuum of the three AFs under investigation. These results are presented in the 'No nil' columns in Table 1. The number of support vectors (SVs) as a percentage of the amount of training data is also listed in Table 1; the values of the γ and c parameters in the SVCs are listed in Table 2 for completeness. The percentage of SVs indicates the SVC complexity: more SVs suggest either more complex decision boundaries or more overlapping data (for an analysis see [4]). The γ is the reciprocal of the RBF kernel width squared and c sets the amount of regularisation.

Table 3 lists the baseline accuracies for each AF value separately and shows that the AF value accuracies differ widely. In the case of 'high-low' and 'fr-back', the accuracy of the "middle" AF value is much lower than those of the "extreme" AF values. The differences between the 'place' AF values are not as large, with the exception of *dental*. For a possible explanation of these low accuracies, see [4].

Table 1. The AF accuracy with (Acc.) and without (No nil) nil and silence frames, and the percentage of the training data that are SVs for the baseline system, EXTR, and REGR.

AF	**BASELINE**			**EXTR**		**REGR**	
	Acc.	*No nil*	*%SV*	*No nil*	*%SV*	*No nil*	*%SV*
'place'	83.1	73.5	40.4	54.4	29.9	58.7	38.2
'high-low'	86.0	69.5	31.0	71.6	23.5	75.1	73.9
'fr-back'	87.1	72.3	40.1	73.2	38.9	73.8	77.0

Table 2. The values of γ, c, and ε (for REGR only) for the baseline system, EXTR, and REGR.

AF	**BASELINE**		**EXTR**		**REGR**		
	γ	c	γ	c	γ	c	ε
'place'	0.1	3	0.01	5	0.01	10	0.5
'high-low'	0.01	3	0.05	5	0.01	1	0.1
'fr-back'	0.01	300	0.1	1	0.1	5	0.1

Table 3. AF value classification accuracies for the baseline system, EXTR, and REGR.

AF	**AF value**	**Acc. (%)**		
		Baseline	*EXTR*	*REGR*
'place'	*bilabial*	70.4	69.2	70.1
	labiodental	71.7	6.3	20.2
	dental	37.1	6.3	19.7
	alveolar	78.8	78.9	78.8
	velar	63.2	36.7	54.9
'high-low'	*high*	74.5	74.6	74.6
	mid	55.2	62.0	72.9
	low	77.6	77.6	77.7
'fr-back'	*front*	83.2	83.3	83.3
	central	35.0	42.6	49.0
	back	61.3	61.4	61.3

4. Modelling the acoustic continuum

4.1. Training on extremes of the continuum ('EXTR')

The success of training SVCs only on the extreme values of an AF continuum and letting the intermediate places of articulation be inferred is dependent on the relative position of each of the AF values in the acoustic (MFCC) space. A "middle" class that is too far removed from the straight line joining the "extreme" classes indicates that it may be less reliable to train SVCs only on the extremes than to train on all three. We therefore calculated the relative positions of the AF value classes.

By inspecting combinations of MFCC coefficients in 3D scatter plots, the distributions of the MFCCs for each AF value class were determined to be "unimodal". The following analysis is performed using 39 dimensional MFCC vectors (without the ±3 frame context window). Disregarding the variances for simplicity, we calculate distances between the mean MFCC vectors of each AF value class and subsequently calculate the perpendicular distance of the "middle" class mean from the straight line joining the "extreme" class means. For an AF consisting of (only) three AF values, it will always be possible to find some sort of continuum through the three class means. The question then is how far the "middle" deviates from the straight line joining the "extremes".

Figure 1 shows 2D representations of the MFCC spaces for the AFs 'high-low' and 'front-back'. Although in both cases the AF value class means do not lie on a straight line the deviation may be sufficiently small that it may be possible to infer a continuum by training SVCs on the extremes, especially for 'high-low'. An investigation of 'place' yielded no obvious straight line in MFCC space that could be used to form a continuum. It suggested that a definition of the 'place' continuum requires a more complex non-linear mapping than can be achieved by simply training on the extremes.

For training the SVC system, only the AF value frames of the "extremes" of the continuums from the 500K training set were used (this method is referred to as 'EXTR'). In the case of 'place', the SVC is trained on *bilabial* and *velar* (30,512 and 40,747 frames respectively). For 'high-low', it is trained on *low* and *high* (56,851 and 61,235 frames respectively). 'fr-back' is trained on *front* and *back* (109,339 and 44,351 respectively). The percentage SVs for each SVC is listed in Table 1; the values for γ and c are listed in Table 2. The SVCs were tested on only the 'relevant' frames in the test material, i.e. the *nil* and *silence* frames were discarded from the test material. This resulted in 139,977 test frames for 'high-low' and 'fr-back' and 199,639 frames for 'place'. The results are presented in terms of 'No nil' accuracy in Table 1.

4.2. Regression ('REGR')

SVR is used to create a function that will be able to describe the articulation continuum (this method is referred to as 'REGR'). Each AF value is assigned a numerical value: for 'high-low', *low*=1, *mid*=2, and *high*=3; 'fr-back', *front*=0, *central*=1, and *back*=2; for 'place', *bilabial*=1, *labiodental*=2, *dental*=3, *alveolar*=4, and *velar*=5. For each AF, SVR is used to fit a function to those values given the 273 dimensional MFCC vector as input. Table 1 shows the percentage SVs and Table 2 shows the values for the parameters γ, c, and ε.

The SVR function is trained on the same data as was used for EXTR but including the frames of the intermediate AF values. So the SVR function is trained on all AF value classes. This resulted in 170,645 training frames for 'high-low' and 'fr-back' and 240,791 training frames for 'place'. The resulting regressions were tested on the same data as EXTR and the 'No nil' baseline. The results are presented in terms of 'No nil' accuracy in Table 1.

4.3. Results

First, we investigate whether the EXTR and REGR classifiers place the AF values in the 'correct' order, i.e. *low-mid-high*, *front-central-back*, and *bilabial-labiodental-dental-alveolar-velar*. To that end, the distributions of the SVR and SVC scores of the test material are plotted. SVR and SVC score distributions show the amount of overlap between any two classes of AF values assuming the 'gold' standard.

The left-hand side of Figure 2 shows the outline of the histograms of the SVR scores of the test material as scored by the REGR (panels a, c, e) and the SVC scores as scored by the EXTR (panels b, d, f) classifiers for 'high-low' (a, b), 'fr-back' (c, d), and 'place' (e, f). As is clear from the SVR score distributions, the REGR classifiers were able to create a function that correctly models the articulation continuum for all three AFs: the distributions of the various AF values are placed in the correct order. In the case of EXTR, for both 'high-low' and 'fr-back' the distribution of the "middle" AF value is clearly placed in between the distributions of the two "extreme" AF values. EXTR is thus able to infer the "middle" AF value after training only on the "extreme" AF values. The EXTR method, however, has slightly more difficulty in modelling the 'place' continuum correctly: *labiodental* and *dental* have swapped places.

In order to compare the results of the REGR and EXTR classifiers with the baseline system, the AF and AF value accuracies need to be calculated by requantising the continuums. To determine the accuracy of any two AF values a threshold can be placed in the SVR and SVC score distributions: every score below the threshold is regarded as one AF value, while every score above it is regarded as the other. ROC curves can be used to plot the accuracies for all possible thresholds. In Figure 2, the right-hand panels show the ROC curves and corresponding score distributions on the left. Comparing the ROC curves for REGR and EXTR shows that REGR seems to outperform EXTR.

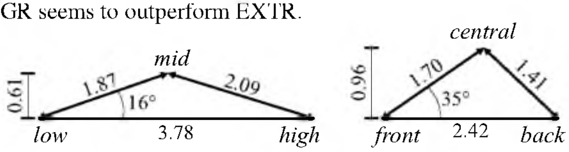

Figure 1. Distances between AF value class means in MFCC space of 'high-low' (left) and 'fr-back' (right).

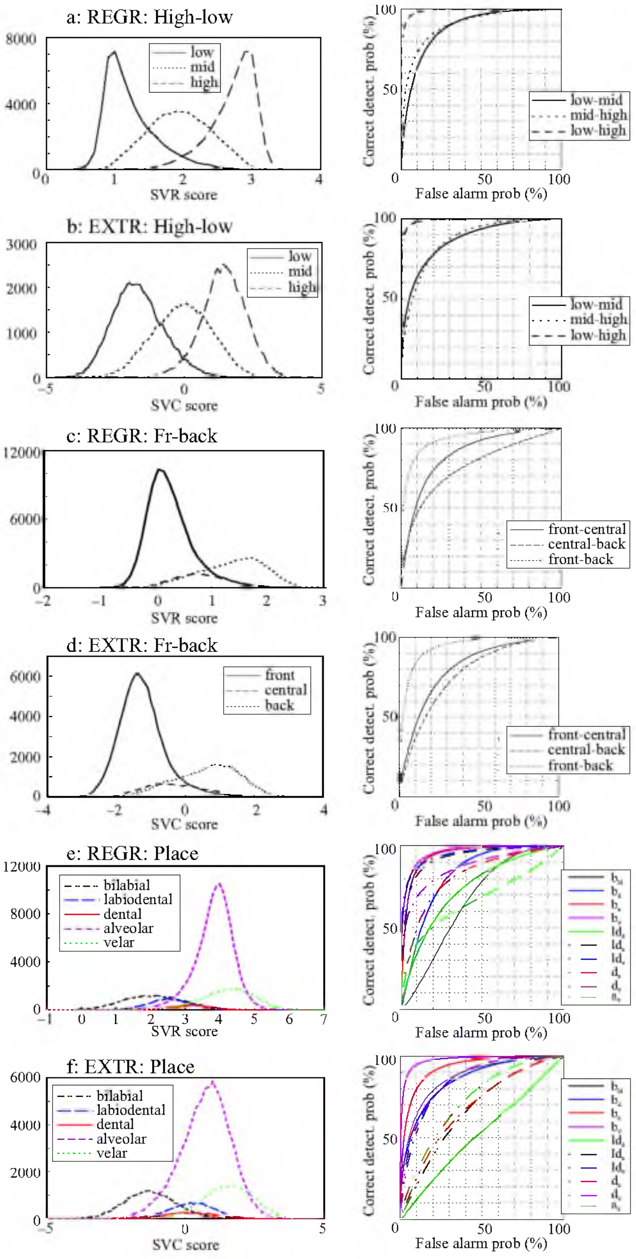

Figure 2. The left-hand side shows the outlines of the test score histograms obtained by the REGR (a, c, e) and EXTR (b, d, f) classifiers for 'high-low'(a, b), 'fr-back' (c, d), and 'place' (e, f); the right-hand side shows the corresponding ROC curves.

The AF and AF value accuracies of the REGR and EXTR classifiers are then inferred by setting thresholds in the SVR and SVC score distributions as described above. The thresholds, for 'high-low' and 'fr-back' are set such that the AF value accuracies for the "extreme" AF values are (almost) identical to the baseline AF value accuracies. For an explanation for 'place', see below. The REGR and EXTR methods are then evaluated on the "middle" AF value(s). Table 1 presents the AF accuracies for EXTR and REGR per AF; the AF value accuracies are shown in Table 3.

A comparison of the baseline system's AF accuracy (Table 1 – 'No nil' column) and AF value accuracy (Table 3) for 'high-low' and 'fr-back' with the EXTR and REGR classifier's accuracies clearly shows what we already hypothesised: modelling of the unquantised articulation continuum results in better AF and AF value accuracies compared to the baseline system. Looking at the AF values shows that EXTR has a 6.8% absolute increase for *mid* and a 7.6% absolute increase for *central* compared to the baseline system; REGR has an even bigger absolute increase of 17.7% for *mid* and a 14.0% increase for *central.*

The results for 'place', however, are different. The thresholds were placed such that the AF accuracies for *bilabial* and *alveolar* match the baseline, since the SVR and SVC score distributions of *labiodental* and *dental* were entirely below the distribution of *alveolar.* This explains the very low AF value accuracies for *labiodental* and *dental.* As Table 3 clearly shows neither REGR nor EXTR were able to model the assumed articulation continuum, in fact they both perform worse than the baseline system.

5. Discussion

In Section 4.1 it was suggested that a definition of the 'place' continuum requires a more complex non-linear mapping than can be achieved by simply training on the extremes. The results in the previous section seem to underline this suggestion. An explanation of the disappointing results might be that for 'place' at least two variables have to be modelled: the *place* of the constriction and the *size* of the constriction. Consider a plosive, e.g. [p], and an approximant, e.g. [w], consonant: the sizes of the constrictions differ hugely with a complete constriction for the first and hardly any constriction for the latter. In the case of 'high-low' and 'fr-back' the size of constriction is fairly constant for all tongue positions. Furthermore, even though the places of constriction for consonants seem to lie on an articulation continuum from the front to the back of the oral cavity, this continuum differs from the 'high-low' and 'fr-back' continuums. During vowel production, the tongue can be at any point between the top and the bottom and the front and the back of the oral cavity. However, during productions of, specifically, *bilabial*, *labiodental*, and *dental* consonants, there is little freedom in the actual place of the constriction; the place of constriction is fixed. These results thus seem to suggest that despite first impressions, there is no real articulation continuum for 'place'. To improve the accuracy for 'place' and its AF values, other methods need to be investigated.

As the results in the previous section show, REGR outperforms EXTR on all levels and for all AFs. An explanation for REGR's better ability of modelling the unquantised articulation continuum is that SVRs provide a natural way to model a continuum and they are trained on all AF values. SVCs on the other hand are binary classifiers trained on only two classes, the "extreme" classes. As hypothesised in Section 4.1, if the "middle" AF value is further removed from the straight line through the "extremes", it may be less reliable to train only on the "extremes". Since SVR is used to create a function to describe the articulation continuum, it does not suffer if the "middle" AF value class is not on a straight line through the "extreme" AF value classes. This hypothesis thus also explains the slightly bigger overall improvement for 'high-low' compared to 'fr-back', since the "middle" AF value class for 'high-low' is closer to the straight line through the "extremes" than for 'fr-back'. The bigger overall improvement for 'high-low' might however also be explained by the fact that for 'high-low' the number of training frames for the "extreme" AF values is approximately equal, while in the case of 'fr-back' there are 2.5 times more *front* than *back* frames. This is a topic for future research.

6. Concluding remarks and future work

The analyses reported here show that the current definition of AFs is not perfect for automatic detection. For 'high-low' and 'fr-back', it is possible to improve on the AF value classification accuracies by modelling the unquantised articulation continuum, but the results from 'place' suggest that new definitions of the AF descriptions and/or alternative modelling approaches may be needed. This is the focus of our future work.

In future work, in our search for a better description of the speech signal, we will also take into account one of the big questions in ASR and psycholinguistics: what is the unit of speech recognition? Psychologists investigating child language acquisition have found that young children learn their mother language by grouping together elements in the speech signal having strong associations with one another, but weak associations with elements within other chunks [10,11]. In analogy with this, we intend to look at methods for unsupervised clustering of speech, e.g. as used by [12], so as to arrive at a method of modelling the unquantised articulation continuum. This approach will, additionally, provide a new and improved definition of the unit of recognition.

7. References

[1] K. Kirchhoff, *Robust speech recognition using articulatory information*, Ph.D. thesis, University of Bielefield, 1999.

[2] S. King, P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language, 14*, 2000, 333-353.

[3] M. Wester, "Syllable classification using articulatory-acoustic features," *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 233-236.

[4] O. Scharenborg, V. Wan, R.K. Moore. "Capturing fine-phonetic variation in speech through automatic classification of articulatory features", *Proc. Workshop on Speech Recognition and Intrinsic Variation*, Toulouse, France, 2006, pp. 77-82.

[5] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," *Proc. IEEE ASRU*, Keystone, CO, 1999, pp. 79-84.

[6] J. Clark, C. Yallop, *An introduction to phonetics and phonology, 2nd edition.* Oxford, UK: Blackwell Publishers Ltd, 1995.

[7] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery, 2 (2)*, 1998, 1-47.

[8] C.-C. Chang, C.-J. Lin, "LIBSVM: a library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[9] J.S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database", *National Institute of Standards and Technology (NIS)*, Gaithersburgh, MD, 1988.

[10] F. Gobet, P.C.R. Lane, S. Croker, P.C-H. Cheng, G. Jones, I. Oliver, J.M. Pine, "Chunking mechanisms in human learning," *TRENDS in Cognitive Sciences, 5 (6)*, 2001, 236-243.

[11] P. Fikkert, "Getting sounds structures in mind. Acquisition bridging linguistics and psychology?" In: A.E. Cutler (Ed.), *Twenty-First Century Psycholinguistics: Four Cornerstones.* Lawrence Erlbaum Associates, 2005, 43-56.

[12] Y. Pereiro Estevan, V. Wan, O. Scharenborg, "Finding maximum margin segments in speech," *Proc. ICASSP*, Honolulu, Hawaii, 2007.