# ARTICULATORY FEATURES FOR EXPRESSIVE SPEECH SYNTHESIS

*Alan W. Black*[1]*, H. Timothy Bunnell*[2]*, Ying Dou*[3]*, Prasanna Kumar Muthukumar*[1]*, Florian Metze*[1]*, Daniel Perry*[4]*, Tim Polzehl*[5]*, Kishore Prahallad*[6]*, Stefan Steidl*[7]*, and Callie Vaughn*[8]

[1]Language Technologies Institute, Carnegie Mellon University; Pittsburgh, PA
[2]Nemours Biomedical Research; Wilmington, DE – [3]Johns Hopkins University; Baltimore, MD
[4]University of California at Los Angeles; Los Angeles, CA
[5]Deutsche Telekom Laboratories/ Technische Universität Berlin; Berlin, Germany
[6]International Institute of Information Technology; Hyderabad, India
[7]International Computer Science Institute; Berkeley, CA – [8]Oberlin College; Oberlin, OH
{awb|fmetze}@cs.cmu.edu

## ABSTRACT

This paper describes some of the results from the project entitled "New Parameterization for Emotional Speech Synthesis" held at the Summer 2011 JHU CLSP workshop. We describe experiments on how to use articulatory features as a meaningful intermediate representation for speech synthesis. This parameterization not only allows us to reproduce natural sounding speech but also allows us to generate stylistically varying speech.

We show methods for deriving articulatory features from speech, predicting articulatory features from text and reconstructing natural sounding speech from the predicted articulatory features. The methods were tested on clean speech databases in English and German, as well as databases of emotionally and personality varying speech.

The resulting speech was evaluated both objectively, using techniques normally used for emotion identification, and subjectively, using crowd-sourcing.

***Index Terms—*** speech synthesis, articulatory features, emotional speech, meta-data extraction, evaluation

## 1. INTRODUCTION

This paper reports results from the "New Parameterization for Emotional Speech Synthesis" group of a workshop which was held at Center for Language and Speech Processing at the Johns Hopkins University in Summer 2011 [1].

Over the last few years, speech synthesis research has moved from using unit selection speech synthesis technology, where subword instances of speech are selected from large databases of natural speech, to a new technology called Statistical Parametric Speech Synthesis (SPSS or HMM synthesis), where generative models of speech are constructed. In spite of the obvious advantages of a generative speech model, the quality of pure statistical parametric speech synthesis systems has not yet surpassed the naturalness of the best unit selection ones, but an explicit model offers a new opportunity for more interesting modeling of speaker specific phenomena.

Much of the work on statistical synthesis uses standard representations of spectral parameters (MFCCs, LSP) but the techniques do not require such a direct parameterization. In this work we are therefore investigating significantly different parameterizations, which we believe are easier to manipulate in order to increase variability in speech synthesis.

In addition to requiring that basic interpolation and distance metrics are well-defined, an ideal parameterization must: (a) be automatically and robustly *derivable* from recorded speech; (b) be able to *reproduce* high quality speech from the parameterization; (c) be able to *predict* the parameterization for text; and (d) capture the *variance* of the speech in the interesting dimensions (speaker identity, emotion, dialect, style, register, etc.)

Potential parameterizations include: **Articulatory Features (AFs)**, i.e. non-segmental features of speech such as nasality, aspiration, voicing, etc. [2]; **Articulatory position data**, as derived from systems like electromagnetic articulograph (EMA) as found in the MOCHA data [3]; and **"Klatt"-like features**, as used in Klatt format synthesis [4] and in KlattStat [5].

In this paper, we will present our specific experiences with Articulatory Features (AFs). During the workshop we also investigated the use of the **Liljencranz-Fant** model, which provides an explicit mode for excitation, but we do not report those results here.

## 2. DATA SETS

In order to investigate how well AFs can represent the variation found in natural speech, we applied our techniques to both standard neutral single-speaker databases in English and German [6], as well as multi- and single-speaker emotion and personality databases:

**LDC Emotional Prosody Speech and Transcripts** (LDC2002S28) contains English dates and numbers from 7 actors: 2418 utterances, average 3 sec, total ≈ 2h. The database contains a 4 class problem (happy, hot-anger, sadness, neutral), a 6 class problem ([...], interest, panic), and a 15 class problem ([...], anxiety, boredom, cold-anger, contempt, despair, disgust, elation, pride, shame).

**Berlin Emotional Database (emoDB)** [7] contains German semantically neutral utterances from 10 actors: 535 utterances, average 2.8 sec, total ≈ 25 min. The 6 emotions are (in addition to neutral): anger, boredom, disgust, anxiety/ fear, happiness, sadness.

**Berlin Personality Database** [8] contains one professional German speaker acting both high and low targets on the "Big Five" personality scales. The database contains 3 parts, with the same and free text spoken in eleven different personalities, ≈ 5 h of speech in total.

For automatic classification of emotions and personality, we extracted 1582 features using openSMILE [9]. We extracted 124 prosodic features (72 $F_0$, 38 energy, 154 duration/ position), 140 voice quality features (68 jitter (JT), 34 shimmer (SH), 38 voicing (VC)), and 1178 spectral features (570 MFCC, 304 MEL, 304 LSP).

## 3. ARTICULATORY FEATURES

In this work, we are not attempting synthesis by inversion. Rather, we view articulatory features as a representation of the intended perception of the speech signal by a listener, similar to own earlier work [2]. Modeling speech using multiple parallel feature streams allows going beyond the "beads-on-a-string" model [10] of speech, and earlier work shows that AFs are well suited for modeling changes in hyper-articulated speech [11], which we regard as a prototype of a strong emotion. Lisping was also found to clearly affect isolated AFs in speaker adaptation [12].

We are therefore not trying to manipulate a physical model of the exact position of the tongue, lips, etc., but we seek to work on a description of the perception of a sound, and hope to be able to show that the observed variations are systematic and meaningful. Generally, we expect that a set of features will generally map 1:1 to speech sounds, even though this is not strictly enforced.

Also, AFs are generally regarded as being dialect and language independent, so that our proposed scheme might be suitable for language-independent or cross-lingual speech synthesis as well.

### 3.1. Generating AFs from Audio

In this work, we will compare three different approaches to including AFs into speech synthesis:

- Purely binary: good for disambiguation, inspired by phonology – containing 40 to 80 binary classifiers [2]
- Multi-stream classification: used for recognition – ca. 8 multi-valued individual classifiers [13]
- Continuous representation – one network, trained to give a continuously-valued vector output, which however is not necessarily a posterior probability

In our experiments, we decided to focus on the third, continuous, representation, for a number of reasons: when trying to predict AFs using Artificial Neural Networks (ANNs), this approach is similar to ASR "bottle-neck" front-end feature representations, which have been shown to be robust against gender and other traits, which we want to normalize. Also, in multi-stream classification, vowels and consonants are treated separately, which opens the question of what to do about semi-vowels, diphthongs, affricates, plosives, voicing? These do not map very well. Our first task will therefore be to compare different AF representations with respect to observe changes between emotions, styles, etc., and investigate their suitability for training, categorization, etc.

As an example of this continuous representation, Figure 1 shows the continuous output of the "place of articulation" node of a neural network trained using QuickNet's[1] "continuous" mode using a 0.4 sec. window of stacked MFCC features as input.

We are currently trying to optimize the prediction of AFs w.r.t. various training error metrics, and learning a topology-preserving mapping as in Figure 1, for comparisons across databases, languages, speaking styles, etc. Similar results have been achieved for other speakers and input representations.

[1] http://www.icsi.berkeley.edu/Speech/qn.html

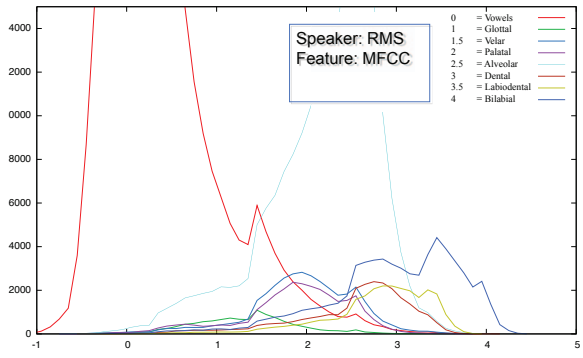


**Fig. 1**. Output distribution (quasi-histogram) of the "front-back" node of an ANN for sounds belonging to different AF categories, trained with the target values shown in the legend. The learned distributions for the 8 classes exhibit inversions of articulatory targets and bi-modal distributions, which, according to manual analysis, mostly stem from improperly labeled, or insufficiently prepared data.

### 3.2. Generating AFs from Text

The AF parameterization is only useful in a text-to-speech environment if it can be predicted from text. We used our standard Clustergen [14] statistical parameteric synthesis system to predict AFs from text. We take the AF predictions from the previous sections at 5 msec intervals and combine these AFs and MCEPs into a supervector. The vectors are then labeled with a large number of contextual features including sub-state position, phone context, syllable context, etc. We build CART trees for each HMM-state labeled set (three per phone). The tree asks context questions and predicts a vector of Gaussians at its leaf. The optimization function for the questions during the building of the CART is minimizing the variance in the AF part of the supervector. This is exactly the same technique we use in building an MCEPs predictor, just in this case we are clustering on the AFs rather than the MCEPs.

To test the effectiveness of such a model, we used the CARTs to predict feature vectors for each frame in a set of held out sentences. We then calculate the Mel Cepstral Distortion (MCD) between the predicted MCEPs and held out set. MCD is a standard measure used in SPSS and Voice Conversion.

We tested on three standard databases: "RMS" (ca. one hour of English male speech), "SLT" (ca. one hour of English female speech) and "FEM" (ca. 30 minutes of German male speech).

In the following, the prediction of MCEPs was done by using the Gaussians of the MCEPs of the features in the leaves of the trees (even though in the AF case the MCEPs were not used directly in the CART question selection). The MCEP example is our baseline using no AFs at all. All examples use Maximum Likelihood Parameter Generation (MLPG, for smoothing MCEP) and the Mel Log Spectrum Analysis (MLSA) filter for re-synthesis. "13c" represents 13 continuous AFs, and "26b" represents 26 binary AFs, as motivated in the previous section.

| | 13c | 26b | MCEP |
|---|---|---|---|
| RMS | 5.360 | 5.320 | 5.197 |
| SLT | 5.284 | 5.278 | 5.140 |
| FEM | 5.822 | 5.761 | 5.600 |

MCD is a distortion measure, so lower is better. A difference of 0.12 is about equivalent to doubling the data, and you probably cannot

hear differences less than 0.07 [15]. Thus the above AF-base synthesis is measurably worse than not using AFs but it is close, and without careful listening tests sounds the same.

As the AFs are predicted without knowledge of their own AF context we added smoothing ("S") to them, and we added AF deltas ("D") to the supervectors. We used a simple 5-point smoother (five times) and added delta features.

| Smooth/ Delta | 13cSD | 26bSD | MCEP |
|---|---|---|---|
| RMS | 5.310 | 5.274 | 5.197 |
| SLT | 5.218 | 5.203 | 5.140 |

This improved the quality, but the AF cases are still not as good as the MCEP alone. The secondary stage we use in Clustergen is to move the HMM-state labels to optimize the prediction quality of the models. Move-label ("ml") is an iterative algorithm [16] that typically improves the MCD score by 0.15 to 0.20. We find:

| Move-Label | 13cSDml | 26bSDml | MCEP |
|---|---|---|---|
| RMS | 5.141 | 5.047 | 5.018 |
| SLT | 4.998 | 4.961 | 4.974 |

Interestingly the move label algorithm gives better gains for the AF based models than the MCEP models. This may be due to the fact that the original boundaries were derived from MFCCs. Now the AF system marginally beats MCEP models for SLT and reaches close in the RMS case. We would not wish to claim the AF models produce better raw synthesis in the case, but do wish to claim the difference between an AF-base system and an MCEP system is negligible.

### 3.3. Mapping AFs to Cepstral Coefficients

The above figures are all based on using the joint MCEPs from the AFs cluster trees. We also investigated building direct models. Using neural nets we trained models for prediction of MCEPs direct from the context of 5 AFs.

For the SLT voice the neural network gave an MCD of 4.97 on the held our test set and 4.91 on the training set, but these AFs were not from our TTS system, but from the original labeling. When put into our TTS system we got 5.45 (as opposed to 5.28 for the joint MCEP prediction). Feeling that there was still something worthwhile in a separate prediction system for MCEPs we investigated an adaptation technique. The AFs we predict with the initial MCEP source are almost certainly noisy. As we are looking for an optimal parameterization that can be predicted by text, and can best produced the desired MCEP we implemented a simple iterative adaptive algorithm. For each set of AF Gaussians in the cluster tree we calculated the error in with respect to the training data. We then adapted AFs to a small percent of that error and retrained the AF to MCEP neural net. We iterated (6 times) until the error ceased to decrease. This system gave an MCD of 5.24.

This technique looks promising though is computationally expensive to train, but the we do not yet know if the move label algorithm is addressing a similar part of the error space. The AF values may not be optimal, so changing them slightly could give a better result, as both this technique, and the above smoothing have done.

## 4. EVALUATION OF EMOTIONAL SYNTHETIC SPEECH

Given the above results on non-emotional data, we look at how best to evaluate our result when using AFs to synthesis more varied speech. We investigated two methods, an objective method based on emotion ID techniques, and for confirmation a subjective method to ensure our objective measures correlate with human perception.

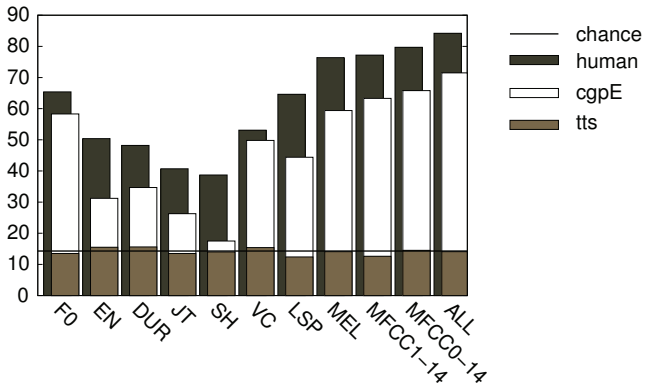


**Fig. 2**. Automatic "objective" classification of emotions on Human and synthetic speech: the "human" bars in the background show UAR (Unweighted Average Recall) on Human speech from emoDB, the horizontal line marks the chance level. The "tts" bars show how these classifiers label *non-emotional*, fully synthesized speech, which is almost at chance level, as expected. The "cgpE" (re-synthesized) speech gets recognized similar to Human speech.

### 4.1. Objective Evaluation

First, we verified that established approaches to automatic detection of emotions in human speech can also be used to detect emotions in synthesized speech of various qualities. Figure 2 shows the unweighted average recalls (UARs) of emoDB emotion classes achieved by various types of features extracted using openSMILE [9] and using WEKA [17] for classification. For the purposes of this paper, we present the following conditions:

**tts** Text to speech *without* emotion content. Predicts durations, $F_0$, and spectrum (through AFs)

**cgpE** text-to-speech with emotion flag, (with natural durations). Predicts $F_0$ and spectrum (through AFs)

We see that automatic emotion classification can be used for synthesis evaluation, and that spectral features are most reliable over all databases (not shown here). We achieve comparable results for English and German, so that the proposed method passed a sanity check for assessing synthesized speech.[2]

Further experiments confirm this impression, and in ongoing work we are investigating the conditions under which certain features (spectral, duration, etc.) can influence the automatic assessment of not the linguistic content of a message, but the perception of the speaking style, in which it is delivered.

### 4.2. Subjective Evaluation

Given the short timeframe available during the workshop, we decided to use crowd-sourcing using Amazon Mechanical Turk (AMT) as our "subjective" verification instrument. We ran a number of verification experiments, to make sure that AMT evaluation produces meaningful results, even if workers may not be using good audio equipment, may be in noisy environments, or may actively try to cheat.

[2]On the LDC Emotion database, this method can predict emotions from the linguistic content (dates & numbers) even if NO emotion parameters are used in synthesis, because certain "non-emotional" words, i.e. years, are not distributed randomly across all emotions.

Almost all workers on AMT speak English, so we first evaluated performance on the English LDC emotion database, using standard and ad-hoc measures to exclude unreliable workers and tasks. Using 74 unique workers, which had completed 169 Human Intelligence Tasks (HITs), we achieved an average classification accuracy of 60% on the four-class problem (anger, sadness, neutral, happiness), which most confusions appearing between happiness, neutral, and sadness. On the fifteen-class problem, we achieved an average of 12% (neutral=29%, hot anger=26%, sadness=25%, ..., anxiety=5%, disgust=5%, shame=4%), with most confusions occurring between sadness, neutral, and contempt (68 workers, 218 HITs).

Using the same setup, the German Berlin Emotion Database's seven-class problem was classified with 41% accuracy, using 37 workers and 245 HITS, which seems reasonable (given that AMT workers are probably not German speakers) and is between the two accuracies achieved for the two conditions of the LDC database. We conclude that AMT can also be used for cross-lingual experiments on emotion recognition, and possibly other voice characteristics.

Taken together, these experiments establish that humans are significantly more accurate than chance for smaller numbers of emotions even in cross-lingual experiments, and with less-controlled settings such as AMT. In our experiments, emotions such as sadness, neutral, and hot-anger could be identified best.

## 5. SUMMARY AND NEXT STEPS

The workshop has established a framework to automatically extract Articulatory Featuress from speech, generate them using text, and synthesize speech using these or the Liljencrants-Fant model, integratint both with automatic and crowd-based evaluation.

Articulatory Features are a competitive and meaningful parameterization of the speech signal, not just arbitrary PCA, which can be independently modified. Neural Networks and CART trees were used to successfully map between spectral features, text, and feature representations on new databases.

The successful use of AFs opens up new research areas concerning more suitable choices of AFs for speech synthesis, more complex generative models, and improvements that can be achieved using for example embedded training. AFs should also be integrated into voice conversion techniques.

Future evaluations of speech synthesis systems should also benefit from our initial results that show that objective and subjective measures of speech qualities are related, also if synthesized speech is evaluated using signal-based measurements of voice quality, and crowd-sourced, cross-lingual evaluations of voice quality. Having an automatic technique available to assess the quality of a voice not only with respect to the linguistic message, but also with respect to a certain emotion or personality will be important for future research.

As part of this workshop, we extracted AF parameters for several databases (ARCTIC, LDC Emotion, Berlin Emotion, Berlin Personality), and developed software which we will add to the FestVox tools distribution. We have also developed the LF-model feature extraction and tools for subjective and objective tools for evaluation of emotional speech synthesis. These are currently being integrated into a single Virtual Machine, which can be run in Virtualbox or other virtualization environments, so that other groups can easily (re-)use our research tools and results.

## 6. ACKNOWLEDGMENTS

This paper is based on work which was conducted by the authors during the 2011 Johns Hopkins University Summer Workshop [1]. This workshop was partly supported by grants from NSF and Google; the authors gratefully acknowledge the faculty, students, and staff at Johns Hopkins who made the summer workshop possible.

## 7. REFERENCES

[1] "New Parameterization for Emotional Speech Synthesis," http://www.clsp.jhu.edu/workshops/ws11/groups/npess, 2011.

[2] F. Metze, "Discriminative speaker adaptation using articulatory features," *Speech Communication*, vol. 49(5), 2007.

[3] A. Wrench, "The MOCHA-TIMIT articulatory database," Queen Margaret University College, 1999.

[4] D. Klatt, "Review of text-to-speech conversion for english," *JASA*, vol. 82, pp. 737–793, 1987.

[5] G. Anumanchipalli, Y. Cheng, J. Fernandez, X. Hunag, Q. Mao, and A. Black, "Klattstat: Knowledge-based parametric speech synthesis," in *Proc. SSW7*, 2007.

[6] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. CMU-LTI-03-177, CMU, Pittsburgh, PA, 2003.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. INTERSPEECH*, Lisbon; Portugal, Sept. 2005, ISCA.

[8] T. Polzehl, S. Möller, and F. Metze, "Modeling speaker personality using voice," in *Proc. INTERSPEECH*, Firenze; Italy, Aug. 2011, ISCA.

[9] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, New York, NY; USA, 2010, MM '10, pp. 1459–1462, ACM.

[10] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. ASRU*, 1999.

[11] H. Soltau, F. Metze, and A. Waibel, "Compensating for hyperarticulation by modeling articulatory properties," in *Proc. ICSLP*, 2002.

[12] F. Metze and A. Waibel, "Using articulatory features for speaker adaptation," in *Proc. ASRU*, 2003.

[13] K. Livescu, M. Cetin, O. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko, "Articulatory feature-based methods acoustic and audio-visual speech recognition," in *Proc. ICASSP*, 2007.

[14] A. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proc. INTERSPEECH*, 2006.

[15] J. Kominek, T. Schultz, and A. Black, "Synthesizer voice quality on new languages calibrated with mel-cepstral distortion," in *Proc. SLTU*, 2008.

[16] A. Black and J. Kominek, "Optimizing segment label boundaries for statistical speech synthesis," in *Proc. ICASSP*, 2009.

[17] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, , and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with java implementations," in *Proc. ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, 1999, pp. 192–196.