New Parameterizations for Emotional Speech Synthesis

Alan W Black awb@cs.cmu.edu, Carnegie Mellon University
H. Timothy Bunnell bunnell@asel.udel.edu, Nemours Biomedical Research
Ying Dou ydou1@jhu.edu, Johns Hopkins University
Prasanna Kumar pmuthuku@cs.cmu.edu, Carnegie Mellon University
Florian Metze fmetze@cs.cmu.edu, Carnegie Mellon University
Daniel Perry djperry@ucla.edu, UCLA
Tim Polzehl tim.polzehl@googlemail.com,
Deutsche Telekom Laboratories/ Technische Universität Berlin
Kishore Prahallad kishore@iiit.ac.in, IIIT Hyderabad
Stefan Steidl, stefan.steidl@informatik.uni-erlangen.de, ICSI
Callie Vaughn, cvaughn@oberlin.edu, Oberlin College

Final Report for NPESS team – CSLP Johns Hopkins Summer Workshop 2011

Abstract

The document gives a description of the work carried out at the 2011 Summer Workshop at CSLP at Johns Hopkins University. This work focuses on finding alternative parameterizations of speech, moving away from more convention spectral representations such as Mel-Frequency Cepstral Coefficients to more speech production related techniques. Specifically we investigated two specific areas. *Articulatory Features*, where the speech signal is represented by multistreams of features that represent phonological features based on IPA-type features, such as place of articulation and/or vowel height, frontness etc. The second area was investigating the *Lilljenkranz-Fant model* where we automatically derive glottal excitation features and formants from databases of natural speech.

As existing statistical parametric speech synthesis techniques already produce fully understandable speech from well-recorded databases, we specifically wished to investigate these two models in an environment that was more demanding than simple read speech. Thus we applied our modeling techniques to databases of varying *emotion and personality.*

Evaluation of speech synthesis is a research field in itself, in that human judgements of quality of synthetic speech are expensive to run and give somewhat subtle results. In this project we have utilized two evaluation techniques, one quite novel, and the other still in very much at an experimental stage. The first novel technique for *objectively* evaluating emotional and personality speech synthesis is to utilized the work on emotion-ID, using such classifiers to score synthetic output. The second *subjective* measure is to use Amazon Mechanical Turk to score the synthesized utterances.

This work has successfully developed novel parameterizations of speech that can produce synthesis of a quality at least equal to existing statistical speech synthesis techniques and shown the effectiveness of using automatic evaluation techniques for wider styles of speech (emotion and personality) than has been done before.

1 Motivation and Background

Over the last 20 year speech synthesis techniques have moved from being an expert-based rule-driven approach to a more data-driven approach. This trend is common in many aspects of speech and language technologies and is due in part to better computing, larger datasets, more demand for faster results, and improvements in modeling and machine learning.

While MITalk [1] best typifies the early work in fully automatic conversion of text-to-speech. Its development required the design and specification of the many stages involved in making natural and understandable speech from text alone. Later work utilized concatenation of human recordings from fixed well-defined databases. At first, databases of all diphones (phone-phone transitions) in a language were carefully recorded and labeled and used to build more natural sounding speech [2]. As inventories became more complex, more elaborate selection techniques were developed to choose the right unit. This led to automatic acoustically based measures to find the right units in a large speech databases [3]. This was later codified into what is now called *unit selection* were selection techniques jointly optimize a target cost and a join cost to find the best units from large databases of natural speech [4].

Unit selection focuses on selection of instances of sub-word (and often sub-phonetic) segments from a databases. Thus the selection techniques can be brittle at the edges of the model, thus badly labeled units can erroneously be selected causing discontinuities in the resulting synthesis that distract the listener.

A further direction in data-driven processing is statistical parametric speech synthesis [5]. Where generative models are used based on averages of speech units. The results are typically much smoother than unit selection techniques, but at a cost of loosing some naturalness or "brightness" to the resulting speech. The annual Blizzard Challenge [6] tests synthesis techniques on the same databases and results show that typical unit selection techniques produce speech that listeners label as more natural, while statistical techniques produce more understandable speech [7].

Much of the current speech synthesis research is focused on the statistical speech synthesis area. Statistical techniques can work well with smaller amounts of data than unit selection, and allow for adaptation techniques that do not fit well into the more traditional unit selection techniques. However like all trends in statistically process there can be over-reliance on abstract subjective measures (e.g. WER in ASR and Bleu in SMT) which can focus research in incremental monotonic improvements. Although this has not yet happened in statistical parametric speech synthesis, the area is mature enough that we understand how well statistical methods work with standard spectral modeling techniques (e.g. MFCCs). But the underlying modeling techniques actually have no need to be restricted to such conventional parameterizations and there is an opportunity to find more relevant representations of speech that better capture how speech is produced, not just how the resulting spectrum is. Many of the parameterizations investigated (often by hand) 20-30 years ago are good representations for speech synthesis but were not practical with that technology.

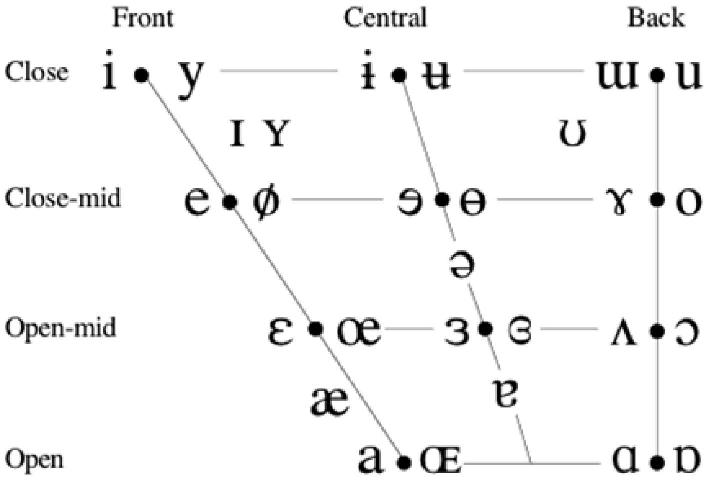
Thus with a goal of broadening the field of statistical parametric speech synthesis into a richer set of representations we proposed this workshop project to look beyond standard spectral modeling.

During the six-week workshop We investigated two different parameterizations in parallel: articulatory features [8] and the Lilljenkranz-Fant model [9]. As SPSS techniques for read speech are quite mature, we also wanted to test these techniques on more varied data thus used databases of emotional speech and varying personalities to see if these new techniques offer a better representation. For evaluation we utilise a novel technique of emotion-ID technology to give an objective measure of our synthesis, and followed this up with a subjective measure using crowdsourcing through Amazon Mechanical Turk.

2 Articulatory Features

2.1 Types of Articulatory Representations

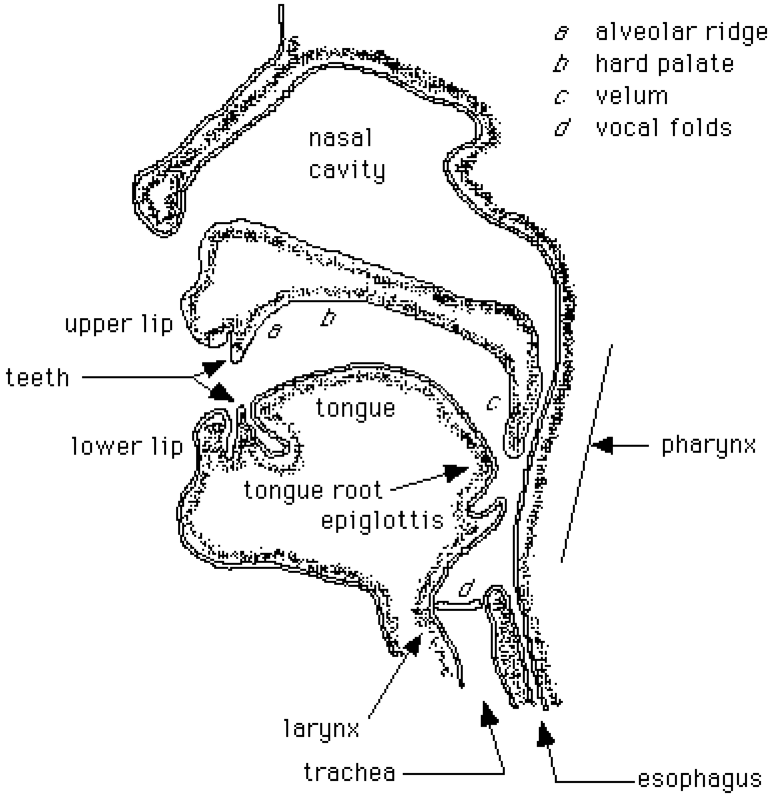
This term has been used to cover a range of related speech representations, thus it is important for us define how we use this term. Traditional phonology as typified by the International Phonetic Alphabet (IPA) identifies phonological segments in speech with specified phonemes that in turn are identified by a collection of features. Vowels in human speech can be viewed in a chart indexed by the frequency of the first two formants.



Where symbols appear in pairs, the one to the right represents a rounded vowel.

We can identify vowels as being in two dimension, high to low (F1), and front to back (F2). Different languages and dialects may make distinctions between different boundaries within this chart, but we can still identify vowels by the amount of heightness and frontness. These are not the only features that identify vowels, roundness, nasality, length, stress and tone may also contribute to phonological variation in articulation.

Consonants too can be broken down in a set of features, distinguishing stops, fricatives, nasals etc. Place of articulation from the lips to the top of the throat.



At this stage we also want to distinguish our use of articulatory features from more explicit representations. In this project we are representing our data as idealized features that we derive from phonological knowledge and directly from acoustics. Other similarly named work uses actual articulatory measurements of human articulators. In such work articulatory position data of tongue, teeth lips, vellum etc of a speaker is measured through techniques such as Electromagnetic Articulatographs (EMA), microbeam or ultrasound. The MOCHA database is one of the common datasets used in such work [10]. Although such work is related to the current project we did not use such data, and have focused on the feature based method described above.

To be even more specific this articulatory feature (AF) techniques follow on directly from the work of Metze [11]. His use of AFs has before concentrated on cross-style speech recognition and in-style recognition but had not yet investigated their use in a synthesis framework.

3 Expressive Speech Databases

In an analogous way that speech recognition has progressed from simple speech styles to more complex ones, (e.g. read speech to conversational speech). Speech Synthesis work has historically concentrated on the synthesis of clearly read speech, typically collected within isolated phonetically rich utterances. Although high quality understandable speech is the result, it is hard to get different styles of speech without also recording data in the target style. Although emotion and style recognition has become a standard field with annual challenges, speech synthesis has only just started looking at expressive databases, and looking for reasonable evaluation paradigms for such generated speech.

As we wished to make our target speech more interesting that just standard read speech we choose to target more expressive databases in order to stretch our techniques and hopefully have a larger space to show off their advantages. We did not wish to record new databases for this work (though in hind-sight that may have helped us).

Due to the availability of databases we choose three different databases of emotional and personality databases. The databases cover both English and German. It seemed useful not to just use one language for our work, but it also should be noted that English and German (at least prosodically and expressively) are relatively similar languages.

LDC Emotional Database

- English, dates and numbers from 7 actors
- 2418 utterances, average length 3s, total about 2 hours
- 4 class problem: happy, hot-anger, sadness, neutral
- 6 class problem: ..., interest, panic
- 15 class problem: ..., anxiety, boredom, cold-anger, contempt despair, disgust, elation, pride and shame

Berlin Emotional Database

- German semantically neutral utterances, 10 actors
- 535 utterances, average length 2.8s, total about 25 minutes
- 6 emotions: anger, boredom, disgust, anxiety/fear, happiness and sadness

In addition to these emotional databases we also made use of a the Berlin Personality database. This data was performed by one actor in ten conditions varying the "Big-5" personality constraints. Specifically there were recordings of plus and minus each of the five emotions: openness, conscientiousness, extraversion, agreeableness and neuroticism.

For each condition there were three types of recording.

Part I The same phonetic recording (around 20s), more than 15 times. Total 960 utterances.

Part II Excepts of 20s from open descriptions of pictures (to provide comparable sized to data to Part I, but more spontaneous speech). Total 210 utterances.

Part III Recording of up to 1.5 minutes of description of pictures. Total 360 utterances (around 5 hours).

All three databases were automatically phonetically labeled using our FestVox tools and complete utterance structure were constructed to allow well defined methods to extract detailed contextual features from them. These databases (and subsets thereof) were used for building synthesizers, voice conversion models, emotion-ID models and for text

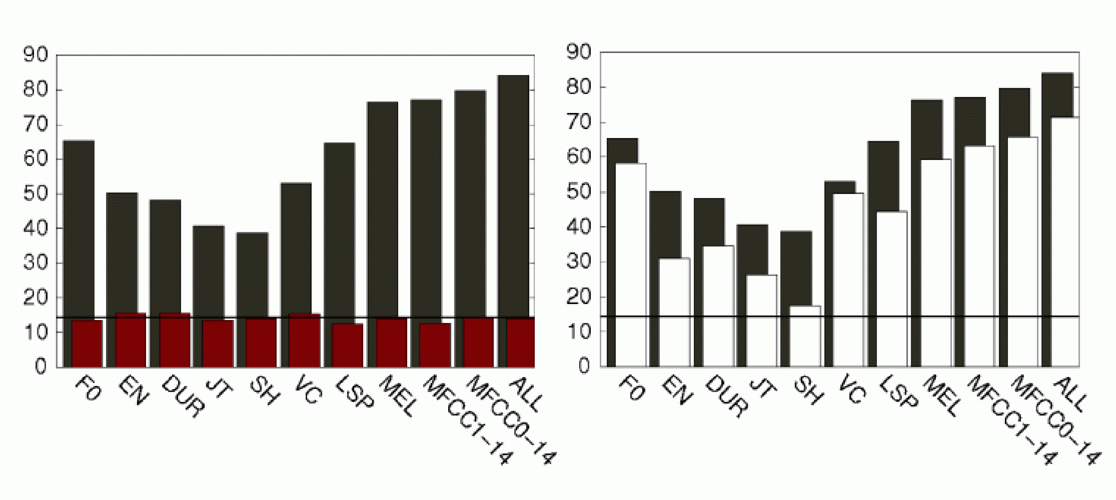


Figure 1: Automatic "objective" classification of emotions on Human and synthetic speech: the "human" bars in the background show UAR (Unweighted Average Recall) on Human speech from emoDB, the horizontal line marks the chance level. The "tts" bars show how these classifiers label *n*on-emotional, fully synthesized speech, which is almost at chance level, as expected. The "cgpE" (re-synthesized) speech gets recognized similar to Human speech.

4 Evaluating Expressive Synthesis

Evaluation of speech synthesis is hard, it fundamentally depends on a listeners personal preference. In order to be able to properly evaluate different models of expressive synthesis it would be necessary to have sufficient listeners in appropriate environments to answer suitable questions about their views. All of conditions are not in themselves well-defined, but we do have some experience on how many listeners are needed, what questions to ask, and how to control for listening conditions. Recently an annual Blizzard Challenge has [6] offered the opportunity to not just allow the testing of different synthesis techniques, but also allow the testing of the stability of evaluation techniques. From its results we can see the types of questions, and the number of listeners we probably require to be confident about significant differences in out models.

Ultimately we want to find an objective measure that is closely correlated with human perception of expressing speech. Such an objective measure could then be used in machine learning optimization off-line. In this project we investigated both objective measures and subject measures.

4.1 Using Emotion-ID systems to evaluate expressive speech

[12] contains much of the description contained here.

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 1 shows the unweighted average recalls (UARs) of emoDB emotion classes achieved by various types of features extracted using openSMILE [13] and using WEKA [14] for classification. For the purposes of this chapter, we present the following conditions:

tts Text to speech *without* any emotion context. 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.[1]

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.

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.

In this chapter, we propose to evaluate the quality of emotional speech synthesis by means of an automatic emotion identification system. We test this approach using five different parametric speech synthesis systems, ranging from plain non-emotional synthesis to full re-synthesis of pre-recorded speech. non-emotional synthesis to resynthesis with all parameters copied from human voices are evaluated. We compare the results achieved with the automatic system to those

[1] 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.

of human perception tests. While preliminary, our results indicate that automatic emotion identification can be used to assess the quality of emotional speech synthesis, potentially replacing time consuming and expensive human perception tests.

4.3 Introduction

In order to improve synthesis of emotional speech, it is necessary to be able to compare different systems and to evaluate their quality. So far, the quality is generally assessed through human perception tests. In order to be able to detect even small differences in the quality of two systems, the number of samples as well as the number of human judges has to be sufficiently high. Finding qualified human participants however is difficult and the number of samples that can be presented to one listener should be limited, to avoid fatigue. Hence, human perception tests are time consuming and expensive. These disadvantages are avoided, if automatic emotion identification could be used as an objective measure to evaluate the quality of emotional speech synthesis. work we explore the usage of automatic emotion identification systems... The underlying assumption is that an emotion synthesis system is of high quality, if the intended emotion can be predicted correctly by an emotion identification system that is trained on human voices. Of course, such measures of emotional quality are meant to complement, not replace, existing evaluation metrics such as Mel-Cepstral Distortion (MCD), Mean Opinion Scores (MOS), or others, focusing on naturalness, intelligibility, or accuracy of the synthesized speech.

4.4 Databases

We used the German "Berlin Database of Emotional Speech" (emoDB) [15] of prototypical emotion portrayals to train and evaluate various emotional speech synthesis systems. 10 actors (5 female and 5 male) produced 10 (emotionally neutral, grammatical, but often non-sensical) sentences each in 7 different emotions: of 7 classes. joy (J), neutral (N), boredom (B), sadness (S), disgust (D), fear (F), and anger (A). For our synthesis experiments, we only retained samples which could be identified with an accuracy of at least 80 % in tests with human listeners. recognized correctly by the human listeners with an accuracy of at least 80 %. Furthermore, the selected samples had to be judged as natural by at least 60 % of the listeners, which leaves 535 samples.

For future experiments, we defined a held-out test set of 69 samples. The classification experiments in this chapter are based on the training set of the remaining 466 samples, using the leave-one-speaker-out evaluation method. same time that samples used for training are disjoint from the ones used for testing. The distribution of the seven emotion categories for both sets is given in Table 1.

4.5 Emotion identification system

Our emotion identification system is based on standard state-of-the-art components: we use the openSMILE toolkit [2] [13] for feature extraction and the WEKA data mining toolkit [3] [14] for classification. We focus on easy to extract acoustic features, and use the 1582 acoustic features of the INTERSPEECH 2010 Paralinguistic Challenge baseline [16]. This feature set is obtained by applying a brute-force approach, in which first of all 38 low-level descriptors and their first

[2] http://opensmile.sourceforge.net/

[3] http://www.cs.waikato.ac.nz/ml/weka/

Table 1: Distribution of the seven emotion categories on training and held-out test set (number of samples). Total amount of speech is ca. 25 minutes, which is acceptable for our purposes.

	training	test	$\sum$
joy	63	8	71
neutral	72	7	79
boredom	73	8	81
sadness	52	10	62
disgust	38	8	46
fear	57	12	69
anger	111	16	127
$\sum$	466	69	535

Table 2: Description of the acoustic features based on 38 low-level descriptors and their first derivative and 21 functionals.

Descriptors	Functionals
PCM loudness	position max./min.
MFCC [0-14]	arithm. mean, std. deviation
log Mel freq. band [0-7]	skewness, kurtosis
LSP frequency [0-7]	lin. regression coeff. 1/2
F0 by sub-harmonic sum.	lin. regression error Q/A
F0 envelope	quartile 1/2/3
voicing probability	quartile range 2−1/3−1/3−2
jitter local	percentile 1/99
jitter DDP	percentile range 99−1
shimmer local	up-level time 75/90

derivative are computed on the frame level. In a second step, 21 functionals are applied in order to obtain a feature vector of constant length for the whole utterance. Table 2 gives an overview of the low-level descriptors and associated functionals. 16 zero-information features (e. g. the minimum of the fundamental frequency is always zero) are removed from the set of 1596 possible features, and two additional features (*F0 number of onsets* and *turn duration*) are added, resulting in a set of 1582 features. As the data consists of emotionally neutral, predefined sentences, no linguistic features are used.

For classification, we used Support Vector Machines (SVMs) with a linear kernel and Sequential Minimal Optimization (SMO) for learning. The complexity parameter was determined in advance and set to 0.1 for the classification experiments reported in Section 4.7.2. As the classes are slightly unbalanced, we applied WEKA's implementation of the Synthetic Minority Oversampling Technique (SMOTE). A 10-fold leave-one-speaker-out evaluation was used to determine the performance of the classifier on the whole data set.

For evaluating a synthesized voice sample, the synthesized voice was treated as additional data of the same speaker as some systems are based on the natural parameters (e. g. natural durations) of this speaker. Thus, neither the synthesized voice nor the data of the corresponding human speaker was seen in the training process. Prior to the classification process, a z-score

Table 3: Different types of acoustic features and number (count) of low-level features derived.

Type	Number
Prosodic Features	
F0	72
energy	38
durations	154
Voice Quality Features	
jitter	68
shimmer	34
voicing probability	38
Spectral Features	
MFCC	570
MEL	304
LSP	304
	1582

speaker normalization of the features was applied.

Since the number of features (1582) is rather high, we also applied principal component analysis (PCA) and feature ranking based on the information gain ratio (IGR) to reduce the number of features. The results show that the SVM classifier can handle the large number of features and that reducing the number of features does not significantly improve the results.

In order to further analyze the differences between synthesized and human emotional speech, we performed separate classification experiments for different sub-sets of features: F0, energy (EN) , duration (DUR) , jitter (JT) , shimmer (SH), voicing probability of the final F0 candidate (VC), line spectral pair frequencies (LSP), logarithmic MEL frequency bands (MEL), and Mel frequency cepstral coefficients (MFCC). different types using the available low-level descriptors as shown in Table 3. Even though our feature set does not explicitly model word or pause durations, the position of the extreme values of all low-level descriptors are durations, and therefore make up a separate group.

4.6 Parametric emotional speech synthesis

For comparison of human and machine evaluation, we created five parametric speech synthesis systems with varying degrees of prediction and hence of different quality. We use "ClusterGen" Parametric Synthesis (CGP) [5], as this will use the data more efficiently than any concatenative technique given the amount of type of training data we have. All systems are based on articulatory features as an intermediate representation [17]. Importantly, we have two dimensions on the systems. The "E" systems (*ttsE* and *cgpE*) include explicit emotion information in the training and testing, i. e. the model uses speech labeled as angry to model angry speech. The non-E systems (*tts* and *cgp*) do not use explicit emotion information, thus acting as controls. The second dimension is changing the amount of information that is predicted, to show the importance of different parts of the signal. The *resynth* system does not predict, but simply decomposes the signal into its components and reconstructs it. *cgpE* and *cgp* use natural durations, and predict spectrum and F0. *ttsE* and *tts* predict F0, spectrum, and durations.

tts Full text-to-speech (TTS) ignoring emotional information in both training and testing. This is a control experiment; the accuracy in the perception and identification experiments should be at chance level (14.3 % for 7 classes).

ttsE As with the *tts* system, this system predicts durations, F0, and spectrum, but also has an "emotion flag" identifying the desired emotion. Training also has this flag, thus the models can generate different emotions. Classification should be better than chance.

cgp As the *tts* system, this system ignores emotions in synthesis and training. It predicts F0 and spectrum, but uses the durations extracted from an original, matching human speech sample. The actual duration patterns are actually dependent on the emotion and – although not modeled explicitly – thus this system actually contains information about the intended emotion of the speaker, and shows the importance of durations.

cgpE As the *cgp* system, this system predicts F0 and spectrum and uses the actual durations from an original speech sample. This system uses emotional labeling in both training and testing, and will generate different predictions for each emotion. We expect the recognition results to be better than the ones for *cgp*.

resynth This system is a pure re-synthesis approach, using natural durations, F0, and spectrum, processed with a speech synthesis framework. It represents an upper limit for the quality of our emotional speech synthesis.

4.7 Evaluation

In order to evaluate the quality of the automatic assessment, the results of the automatic emotion identification system are compared to the ones obtained in a human perception test.

4.7.1 Human perception tests

As human perception tests are time consuming and expensive, we selected an emoDB subset that contains 5 randomly selected samples for each emotion. For each of the 6 experiments, each of the 35 audio samples was presented to 15 native German listeners in random order using a web interface. For each sample the human judges had to select one of the 7 given emotions, resulting in 525 judgments for each experiment. The judges were mostly students (36% male / 64% female, mean age 26 years, age range 22-39) and wore high-quality headphones, in a quiet office environment. Listening and judging took 9.5 seconds on average per sample.

As expected, the results of the *tts* control experiment (15.6 %) are close to chance level (14.3 %). According to the human judges, there is no significant difference between the *ttsE* and *tts* systems, even though *ttsE* includes emotion information. However, if natural durations are used instead of predicted ones, without an emotion flag (system *cgp*), human listeners are clearly able to distinguish the seven emotions. The accuracy for *cgp* is 61.5 %. Again, adding an emotion flag does not lead to better results in the human perception test, in fact leading to an insignificant degradation for *cgpE* (61.0 % vs. *cgp*'s 61.5 %). In our upper limit experiment – the re-synthesis based on natural durations, F0, and spectrum – the human judges can predict the seven emotions with an accuracy of 79.8 %. This is certainly a good result, but it is still worse than the performance of the human listeners for the original recordings of the actors, which is 87.7 %. The accuracies are summarized in Table 4. The confusion matrices are shown in Table 5.

Table 4: Results of the human perception tests compared to the results of the emotion identification system for different synthesized voices and the original human voices.

	Human Perception	Emotion ID	
emoDB	subset	subset	full
tts	15.6 %	14.2 %	14.1 %
ttsE	17.5 %	17.1 %	29.0 %
cgp	61.5 %	62.8 %	64.5 %
cgpE	61.0 %	74.2 %	71.5 %
resynth	79.8 %	85.7 %	81.8 %
original	87.7 %	82.8 %	83.7 %

		hypothesis							
		J	N	B	S	D	F	A	∑
reference	J	**4**	33	1	8	4	16	9	75
	N	4	**46**	5	6	3	7	4	75
	B	0	48	**1**	2	3	18	3	75
	S	5	46	4	**2**	2	12	4	75
	D	1	34	5	9	**4**	17	5	75
	F	11	36	0	2	4	**19**	3	75
	A	0	41	2	3	7	16	**6**	75
									525

(a) tts: **15.6 %** accuracy

	hypothesis						
	J	N	B	S	D	F	A
J	**4**	19	2	6	1	42	1
N	15	**28**	3	16	2	9	2
B	1	39	**3**	15	3	12	2
S	4	24	1	**21**	4	19	2
D	1	31	2	16	**5**	18	2
F	8	29	2	13	3	**20**	0
A	3	19	0	6	5	31	**11**

(b) ttsE: **17.5 %** accuracy

	hypothesis						
	J	N	B	S	D	F	A
J	**40**	18	4	0	1	4	8
N	0	**54**	9	3	0	4	5
B	3	6	**63**	1	1	1	0
S	0	8	16	**47**	1	3	0
D	1	6	3	18	**34**	13	0
F	6	11	0	1	4	**37**	16
A	17	7	1	0	0	2	**48**

(c) cgp: **61.5 %** accuracy

		hypothesis							
		J	N	B	S	D	F	A	∑
reference	J	**41**	19	0	1	2	4	8	75
	N	1	**45**	8	7	3	1	10	75
	B	0	6	**67**	2	0	0	0	75
	S	0	6	25	**40**	1	3	0	75
	D	0	5	1	17	**35**	17	0	75
	F	4	15	4	1	0	**44**	7	75
	A	21	6	0	0	0	0	**48**	75
									525

(d) cgpE: **61.0 %** accuracy

	hypothesis						
	J	N	B	S	D	F	A
J	**63**	6	0	0	1	1	4
N	0	**53**	12	7	0	0	3
B	0	0	**68**	5	2	0	0
S	0	1	8	**63**	0	3	0
D	0	2	0	12	**61**	0	0
F	15	0	0	0	0	**46**	14
A	8	2	0	0	0	0	**65**

(e) resynth: **79.8 %** accuracy

	hypothesis						
	J	N	B	S	D	F	A
J	**69**	1	0	2	1	0	2
N	0	**64**	5	2	1	0	3
B	0	1	**74**	0	0	0	0
S	0	0	5	**69**	0	1	0
D	0	1	2	9	**62**	1	0
F	16	1	0	0	0	**54**	4
A	5	1	0	0	0	0	**69**

(f) original: **87.8 %** accuracy

Table 5: Results of the **human perception tests** for different synthesized voices and the original intended emotions. Using natural durations seems to be important for classification (5 audio files for each of 7 classes, annotated by 15 labelers each = 522 comparisons).

There appears to be no generalizable systematic effect of the *E* emotion flag. Anger and sadness recognition clearly benefits in the *tts* systems. While fear recognition suffers, all other emotions remain near the baseline. When applied to *cgp* systems, the flag inclusion boosts the recognition of all emotions except sadness. Also, accuracy of neutral speech recognition decreases. Consequently, learning durations, F0 and spectral parameters from emotion-specific data partitions generally improves recognition of synthesized anger.

4.7.2 Automatic evaluation based on emotion identification

The emoDB-trained emotion identification system described in Section 4.5 is now used to evaluate the five systems for synthesis of emotional speech described in Section 4.6.

For the three systems *tts*, *ttsE*, and *cgp*, the results of the automatic system on the subset are

		hypothesis							
		J	N	B	S	D	F	A	∑
reference	J	**15**	8	12	0	6	16	6	63
	N	20	**9**	15	0	3	17	8	72
	B	19	9	**13**	0	7	20	5	73
	S	13	10	13	**0**	2	14	0	52
	D	14	4	5	0	**4**	7	4	38
	F	15	8	9	0	5	**15**	5	57
	A	31	16	23	0	7	25	**9**	111
									466

(a) tts: **14.1 %** UAR, 13.9 % WAR

	hypothesis						
	J	N	B	S	D	F	A
J	**33**	5	3	0	0	15	7
N	3	**33**	11	0	19	5	1
B	2	36	**9**	0	18	7	1
S	1	20	12	**0**	16	3	0
D	1	13	4	0	**9**	11	0
F	6	18	3	0	15	**11**	4
A	40	0	0	0	0	16	**55**

(b) ttsE: **29.0 %** UAR, 32.1 % WAR

	hypothesis						
	J	N	B	S	D	F	A
J	**43**	4	0	0	0	10	6
N	0	**57**	4	0	1	10	0
B	0	18	**43**	2	9	1	0
S	0	3	10	**33**	6	0	0
D	2	4	3	2	**25**	1	1
F	3	7	3	0	1	**41**	2
A	51	1	0	0	4	6	**49**

(c) cgp: **64.5 %** UAR, 62.4 % WAR

		hypothesis							
		J	N	B	S	D	F	A	∑
reference	J	**47**	3	0	0	0	7	6	63
	N	0	**61**	7	0	1	3	0	72
	B	0	16	**48**	1	5	3	0	73
	S	0	0	10	**39**	2	1	0	52
	D	4	3	2	2	**25**	2	0	38
	F	2	6	2	0	2	**43**	2	57
	A	31	0	0	0	5	9	**66**	111
									466

(d) cgpE: **71.5 %** UAR, 70.6 % WAR

	hypothesis						
	J	N	B	S	D	F	A
J	**44**	1	0	0	1	8	9
N	0	**62**	5	0	1	4	0
B	0	3	**65**	3	2	0	0
S	0	1	5	**46**	0	0	0
D	0	0	2	0	**32**	2	2
F	6	3	0	1	0	**45**	2
A	23	0	0	0	0	3	**85**

(e) resynth: **81.8 %** UAR, 81.3 % WAR

	hypothesis						
	J	N	B	S	D	F	A
J	**40**	0	0	0	1	7	15
N	0	**64**	6	0	2	0	0
B	0	4	**66**	2	1	0	0
S	0	0	1	**51**	0	0	0
D	1	4	1	1	**29**	1	1
F	4	2	0	1	1	**47**	2
A	13	1	0	0	0	1	**96**

(f) original: **83.7 %** UAR, 84.3 % WAR

Table 6: Confusion matrices and performance of the **automatic emotion identification system** for different synthesized voices and the original human voices in terms of the (unweighted) average recall (UAR) and the weighted average recall (WAR) / accuracy.

very close to the results of the human perception test. For *cgpE* and *resynth*, better results are obtained with the objective measure, whereas the results are worse for the original human voices. However, this subset of 35 samples is very small and hence the significance of these differences is low. For the performance of the emotion identification system on the whole training set, similar trends can be observed. On the whole training set (which, again, we use for leave-one-speaker-out testing), the system *ttsE* is judged clearly better than *tts*, and *cgpE* is clearly better than *cgp*. *Resynth* represents the best of the five speech synthesis systems, still performing slightly lower than the original human voices, though. Table 6 shows the confusion matrices.

Figure 2 shows the performance of emotion identification systems trained on different subsets (or types) of acoustic features. In general, the classifiers based on spectral features (MEL, MFCC 0-14, MFCC 1-14) as well as LSP perform very well. They also contain the highest numbers of individual features, which can bias the evaluation. Inclusion of the emotion flag improves synthesis and objective evaluation on basis of these features, as expected. The same can be observed for F0 features. All other features change inconsistently with respect to the switch. The lowest performance is obtained with the small group of jitter and shimmer features. Evaluation based on VC or F0 features only leads to inconclusive results, as our classifiers seem to detect synthesis predictions of pitch and voicing better than actual resynthesized pitch and voicing.

Table 2) is clearly lower than the number of spectral features. The performance of the *resynth* MEL and MFCC features is almost identical to the performance on human voices. However, a clear drop in performance is observed for F0 features and the voicing probability features of the final F0 candidate. Obviously, there are clear differences between the energy patterns and smaller differences between the durations patterns, too.

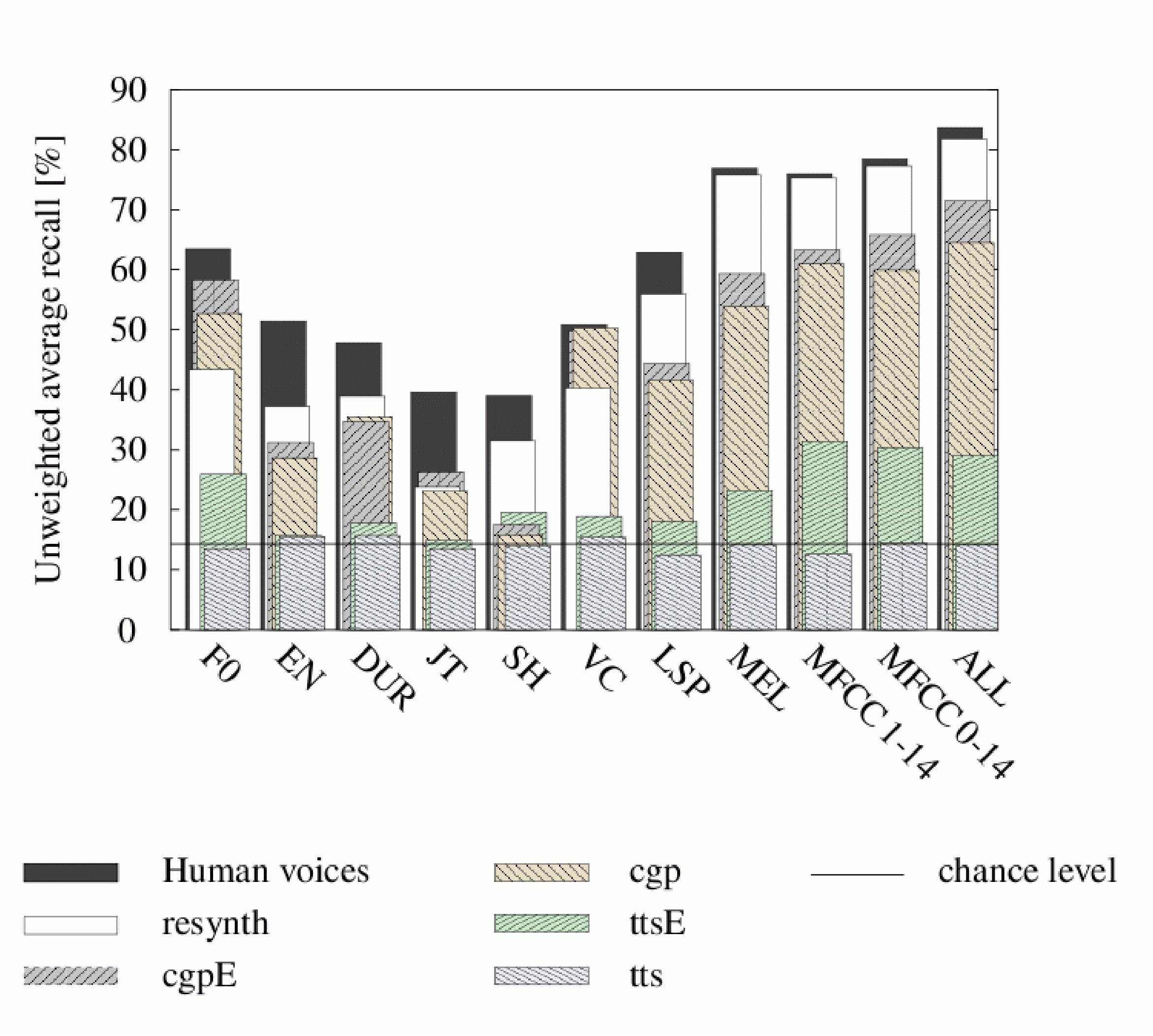


Figure 2: Automatic emotion identification results for different feature types, see Table 2 and Table 3.

4.8 Conclusions

The results of the emotion identification experiments are very consistent and mostly confirm our intuition. The results of the objective measure highly correlate with the ones of the human perception tests – however at a much lower price, much faster, and with much lower effort involved in the evaluation. It is interesting to note that for both human evaluation and evaluation by automatic classification using natural durations seems to be the most important factor to achieve high accuracy.

Thus, automatic emotion identification can be used successfully to judge the quality of emotional speech synthesis systems, at least for in-development assessment of improvements, if not for final judgments. In addition, the analysis of different feature types can give valuable insights into why synthesis systems perform differently, and worse than human voices.

5 LF-Model and Formant Synthesis

Unit selection synthesis [18, 19], based on waveform concatenation, is capable of producing synthetic speech that is almost indistinguishable from natural speech in some circumstances. However, it has a number of widely recognized drawbacks as well. Foremost among these is the difficulty of rendering natural expressiveness in the output. Because the unit selection approach requires every possible variation of speech quality to be represented in the database, a database containing all expressive variations—happy, sad, excited, angry, sarcastic, doubtful, etc.—would be prohibitively large.

Parametric speech synthesis [20, 21] has the potential to avoid this problem since the acoustic speech signal is not stored, but rather generated de novo based on models of the time course of a relatively small number of acoustic phonetic parameters. Thus, any acoustic variation that can be modeled can be generated. However, the problem for parametric synthesis is to develop models of sufficient depth and complexity to capture all the qualities of natural speech.

Recent advances in speech synthesis (e.g., [5]) have led to the development of statistical parametric speech synthesis (SPSS) techniques that learn the time course of parameters and thus do an excellent job of fitting a parametric model to the speech of an individual, capturing both the segmental and prosodic patterns common to that individual. The NPESS project was intended to extend this approach to also learning the underlying patterns associated with different expressive features in speech.

One area in which SPSS has been somewhat weak has been in the area of voice quality. Many of the present SPSS systems have a buzzy quality that has been attributed to the overly simple pulse excitation model that underlies the waveform generation process. Voice quality has also be identified as one of the most important contributors to expressiveness in natural speech (e.g., [22]). For both these reasons, we chose to concentrate on applying a better voicing source model–the Liljencrants-Fant (LF) model [9]–within the context of an SPSS model.

More specifically, this portion of the NPESS project examined the possibility of conjointly fitting formant synthesis parameters (pole and zero frequencies and bandwidths) and source characteristics to speech corpora. We sought to achieve several goals:

1. Update and adapt existing speech analysis software to provide highly accurate paramaterizations of an individual's speech using an analysis-by-synthesis technique.

2. Show that features derived in this way, notably the phonatory source features, were able to represent expressive variation in a way that could be used to both classify and manipulate the expressive quality of the speech.

3. Show that the parameterizations was amenable to current statistical learning techniques and thus, that it could be used to create a parametric speech synthesizer.

We describe progress toward each of these objectives in the following.

5.1 Software

Software adapted for the workshop was based on an analysis-by-synthesis approach described by [23]. The synthesis component of this software was a cascade formant synthesizer based loosely on the design first described by [21], but without a parallel resonator branch. The cascaded formant resonators and zeros were driven by the the Liljencrants-Fant source model for voiced

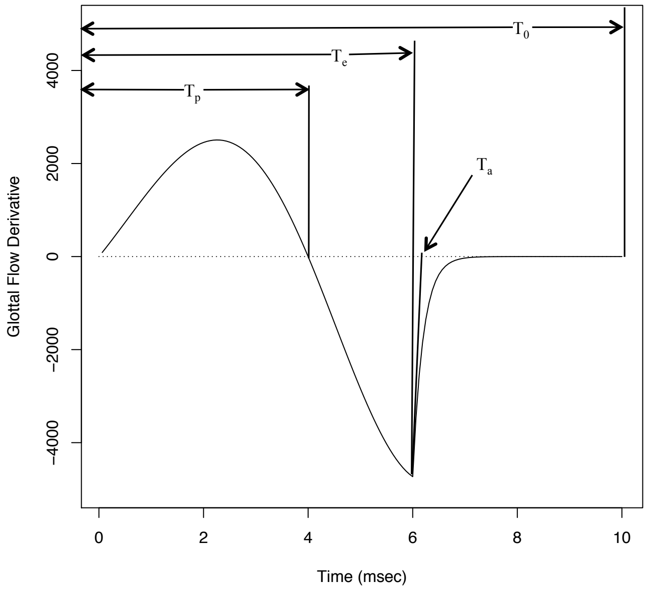


Figure 3: Illustration of the Time-related parameters of the LF model

segments, and a uniform distribution noise generator for voiceless segments. These components are described more fully below. All software developed for the workshop as well as a number of previously developed support libraries and applications will be made freely available under an MIT-like open source license.

5.1.1 Liljencrants-Fant (LF) source model

The LF model [9] describes the glottal source characteristics for voiced speech segments. With relatively few parameters, it is intended to capture the important features of variation in voice quality that have been shown to be important in modeling the expressive aspects of speech (e.g., [24, 25, 22]). The most important shaping parameters of the model are illustrated in Figure 1. The parameters and their definitions are:

$\mathbf{T}_0$ The fundamental period

$\mathbf{T}_e$ The duration from the start of a glottal cycle to the point of maximum rate of closure (a minima in the first derivative)

$\mathbf{T}_p$ The time from the start of a glottal cycle to the point of maximum air flow (a zero in the first derivative)

$\mathbf{T}_a$ The time from Te to the projection onto the 0 flow derivative of the steepest slope in the return from Te.

The implementation of the LF model used for the present project is a translation to C from the Fortran code initially published in Q. Lin's dissertation. The standard LF model was modified slightly to allow a noise source to be introduced during the open phase of each glottal cycle, mimicking the effects of aspiration noise generated in the glottis.

To aid in developing figures and for testing purposes, the LF model was also implemented in R.

5.1.2 Cascade formant synthesizer

The cascade resonator formant synthesis system used for this study is designed to operate as a pitch synchronous system rather than a frame-based system as is more typical. Although written in C, the resonator equations are those used by Klatt (1980). The vocal tract response of the present synthesizer is characterized by the frequencies and bandwidths of up to 10 resonators and 3 anti-resonators. Source models for the synthesizer consist of, for voiced epochs, the LF model augmented by and aspiration noise source that can be introduced during the open phase of the the glottal cycle, and a frication source for voiceless periods.

Table I lists the input parameters to this synthesis program. In Table I, parameters shown in red are alternative higher-order parameters that can be used to generate the basic parameters (shown in black print) for some synthesizer configurations. Briefly, the higher-order source parameters were open quotient (OQ), which was defined operationally as the ratio of Te to T0, and something similar to a speed quotient termed PE, which was the ratio of Tp to Te. Both of these parameters were censored to a reasonable range of values. The Ta parameter was directly fitted in all configurations, but when fitting the higher order parameters OQ and PE, Ta was expressed as a percentage of T0 and constrained to a maximum value < 1.0 - OQ.

The vocal tract-related higher order parameters were based on a quasi-articulatory model and include an overall vocal tract length scaling factor (SCL), which scales neutral formant frequencies by a constant multiplier to model changes in vocal tract length. For this parameter, a neutral value (SCL == 0.0) corresponded to a vocal tract length of about 17 cm and gave rise to formant frequencies starting at 500 Hz and separated by 1000 Hz. Values of SCL < 0.0 correspond to longer vocal tracts (with formants more closely spaced), and values > 0.0 correspond to shorter vocal tracts (with formants more widely spaced). A resonance factor (RES) determines formant bandwidths relative to average typical values for each formant. Values of RES > 0.0 lead to narrower formant bandwidths, while values < 0.0 lead to wider formant bandwidths. A formant frequency shift factor (SFT) raises or lowers all formant frequencies by the same absolute amount, mimicking closure versus flare at the lips. A high-front to low back constriction factor (FBF) mimics the effects of constriction within the entire front half versus back half of the vocal tract. Values of FBF > 0.0 mimic a constriction within the front half of the vocal tract that lowers the frequency of all odd-numbered formants while raising the frequency of all even numbered formants (producing an /i/-like acoustic pattern), and values < 0.0 mimic the opposite pattern in which odd-numbered formants increase in frequency while even-numbered formants decrease in frequency. Finally, a retroflexion/palatalization factor (RPF) was introduced for which values > 0.0 cause F2 and F3 to drop toward F1 (and /r/-like pattern) while values < 0.0 cause F2 and F3 to rise toward F4 (an /y/-like pattern).

As one element of the studies conducted during the workshop, we examined whether it would be better to adapt these higher-order parameters or the raw synthesizer parameters.

Table 7: Cascade formant synthesis parameter sets.

Parameter	Description
AV	Amplitude of Voicing
AH	Amplitude of Hiss
AF	Amplitude of Frication
TA	LF Ta parameter $(0.0 < \text{Ta} < (1.0 - \text{OQ}))$
TE	LF Te parameter
TP	LF Tp parameter
OQ	Open Quotient $(0.2 < \text{OQ}=\text{Te}/\text{T0} < 0.8)$
PE	Tp to Te Ratio $(0.5 < \text{PE}=\text{Tp}/\text{Te} < 0.99)$
F1 - FA	Frequencies for 10 formants
B1 - BA	Bandwidths for 10 formants
Z1 - Z3	Frequencies for 3 zeros
ZB1 - ZB3	Bandwidths for 3 zeros
SCL	VT Length Scale Factor (1.0 = 17 cm)
RES	VT Resonance
SFT	Formant Frequency Shift factor
FBF	High Front to Low Back scale factor
RPF	Retroflextion to Palatalization factor

5.1.3 Analysis by Synthesis (AbS) approach

The AbS approach operates pitch synchronously to fit synthesis parameters to successive voiced or voiceless epochs. Voiced epochs constitute single cosine-windowed pitch periods centered on the point corresponding to Te in the LF model. Voiceless epochs were generally pitch period sized regions of signal excised at arbitrary successive points in voiceless regions and also windowed with a raised cosine window. All parameter fitting takes place in the spectral domain. Because each epoch is at most a single windowed pitch period, harmonic structure is not represented in the spectral domain and the speech spectrum resembles the spectrum of the vocal tract impulse response convolved with the source function. Thus, it is possible to directly compare the magnitude spectrum of the output of the cascade formant synthesizer as excited by an appropriate source function with the magnitude spectrum of the natural speech epoch.

A Levenberg-Marquardt minimization technique is used to adjust the initial guess for all synthesis parameters to minimize the difference between the model and target natural speech spectrum of each frame in terms of a Chi-squared statistic. For details of the L-M minimization approach, see, for example, Press, et al. (1988). One of the issues examined during the workshop was the nature of the initial guess given to the L-M minimization algorithm. It is possible to start the analysis of each epoch independently by setting all parameters to "neutral" values and then allow the algorithm to proceed to a stopping point (where the Chi-square statistic is not being significantly improved), or, to start each new epoch with parameter values set to the best fit from the previous frame. The former approach should minimize the chance of propagating errors due to an ill-fit frame, while the latter should reduce computation time and improve the frame-to-frame consistency of the parameter values obtained.

5.1.4 Waveform display with model fitting module

An MS Windows-based waveform display and editing program (Wedw) was initially used to allow interactive testing of the fitting process. With Wedw, it was possible to select individual epochs to fit, or regions of a file, or an entire file. While the fitting process was running, a spectral cross section display shows the relationship between the model and target spectra so that one can quickly and directly observe how well the L-M algorithm is fitting parameter values. Wedw also allows the user to selectively enable/disable parameters from the fitting process, and to select initial values for each parameter to test effects of changes in the initial values.

This program made it possible to try many different variations of the AbS procedure and monitor the results of the changes. However, Wedw did not provide any method of batch processing multiple waveform files as needed to process entire speech corpora. For that, a non-interactive command line program was developed that used the same underlying fitting procedure as Wedw, but applied in a batch-oriented manner.

5.1.5 Standalone analysis program

The standalone analysis by synthesis program (called *abs*) went through several development stages during the workshop to allow testing various configurations for model fitting. Results of these tests are described below.

5.2 Stimuli

5.2.1 Normal Speech Database

The normal speech database used for testing the LF & Formant parameter extraction was the 'A' set of sentences from the CMU Arctic SLT corpus [26].

5.2.2 Emotional Speech

Analyses of parameter extraction for emotional speech used the CL talker from the LDC emotional speech database. Additionally, we restricted our analyses to just files representing the four emotional states: normal, hot-anger, happy, and sad.

5.3 Procedure

5.3.1 Software development and testing

Throughout the workshop, modifications were made to the underlying library of functions used by both Wedw and the command line program abs to improve the robustness and accuracy of both the formant and LF parameter estimates. As we were attempting to very rapidly make changes to the software, much of the work involved subjective listening to copy synthesis output. However, several formal evaluations were conducted as well to track improvements quantitatively and explore alternative approaches. For this, we looked at both the copy synthesis performance of the algorithms, and also the usefulness of the derived parameters for SPS within the ClusterGen framework [27]. For both types of analyses, the criterion measure was the Mel Cepstral Distance (MCD) between an analysis of an original (natural) utterance and a synthesized version of the same utterance. For copy synthesis, the natural comparison was, of course, the utterance from which parameters were extracted. For SPS evaluations, the comparison utterances were always from a set of utterances that were not used as part of the training corpus.

For discussion, we refer to different versions of the AbS software as:

- Wedw – Wedw-based software that was used initially as the algorithm was refined.
- ABS v1.0 – The first of the batch-processing versions that ran as a command line tool. It was a fairly direct implementation of the Wedw-based process, with a few bug fixes.
- ABS v2.0 – A version in which many of the functions used by Wedw were rewritten to work more efficiently in a stand-alone environment. The process as implemented within Wedw required much of that program's infrastructure to operate and replicating the infrastructure was cumbersome for a program that was not intended to be used interactively. ABS v2.0 was our first attempt to replicate the same functionality in a way that was better suited for a command line tool.
- ABS v2.1 – This was the version of the ABS program used to examine the issue of independent versus dependent parameter initialization, and also was used to compare fitting the quasi-articulatory parameters versus raw feature values. Most of the changes from v2.0 were either bug fixes or feature additions that did not greatly change the underlying approach of v2.0.
- ABS v2.2 – With this version an attempt was made to employ a more global fitting process that, rather than fitting each frame until parameters had converged on a final solution, made several passes over the entire collection of frames for an utterance to find regions where parameters seemed to be independently converging on a similar solution, then constraining the parameter fitting within those regions. It was thought that this might lead to a "best of both worlds" solution that both encouraged consistent frame-to-frame fits, but did not tend to propagate fitting errors due to a single poorly fit frame processed in sequence.

5.3.2 LF model parameters and emotional speech

In addition to the effort to improve parameter fitting (our major effort during the workshop), we were also concerned that the parameter we obtained for the LF model did in fact reflect changes in expressive speech states. To establish this, we analyzed the speech of one talker (CL) from the LDC emotional speech corpus with four types of emotional speech: *neutral, sad, happy*, and *hot-anger* (hereafter, *angry*). For each utterance, we averaged the values of AV, T0, Tp, Te, and Ta over the entire utterance to obtain an average value of each example utterance. There were multiple utterances by this talker representative of each of the four emotions. We then treated each utterance as a sampling unit for purposes of analysis of variance to determine if there were significant differences in the parameter values associated with sentences uttered with different acted emotions. Separate analyses of variance were run for each of the parameters of interest.

As an additional check on whether the parameter values were predictive of the underlying utterance emotion, we trained linear discriminant functions to predict/classify emotion based on the parameters and then tested the accuracy of the prediction/classification via a leave-one-out cross validation study.

Table 8: Mel Cepstral Distances for copy synthesis and statistical parametric synthesis.

Analysis Method	Copy Synthesis MCD	SPSS MCD
Wedw	7.6	8.07
ABS v1.0	6.87	7.97
ABS v2.0	_	7.28
ABS v2.1	5.66	6.68
ABS v2.2	6.17	6.72

Table 9: Mel Cepstral Distances for copy synthesis based on two parameter sets (Articulatory versus Raw) and two methods of initializing the parameter search for each frame (Dependent upon the previous frame results versus Independent of the previous frame and starting from a neutral VT configuration.)

Condition	MCD
Artic/Dep	6.95
Artic/Ind	6.84
Raw/Dep	6.60
Raw/Ind	6.41

5.4 Software development and testing

Results of our several formal tests of software versions are shown in Table 2 in terms of the MCD for copy synthesis tokens, and for the results of training a ClusterGen system with the obtained parameters. As this table suggests, performance generally improved over the course of the workshop; Starting from the original Wedw version, MCD decreased for most subsequent versions except the final version which tended to yield somewhat larger MCDs. Data from the copy synthesis for the v2.0 program are not reported because there were a large number of instances where the algorithm apparently failed to converge on valid parameter settings. This led to a significantly bimodal distribution of tokens with some sounding good and having lower MCD scores while others were virtually unintelligible and yielded high MCD scores; the mean of such a bimodal distribution is essentially meaningless. The ClusterGen analysis of those stimuli do not reflect this problem, presumably because the ClusterGen results are based on parameter distributions derived from a much larger number of analyzed tokens. That is, copy synthesis results are based on analysis of 9 individual files, while ClusterGen results are derived from mean and standard deviations of about 500 analyzed files.

Using the 2.1 version of ABS, we also examined alternative ways of fitting parameters and the possible use of an alternative (quasi-articulatory) parameter set. Results of this are shown in Table 2, which shows that MCDs were greater for the quasi-articulatory parameters than for the raw parameters, and greater for analyses in which the frames were fit dependent upon the previous frame versus independently starting from neutral values on each frame.

Table 10: Average LF parameters associated with each of four emotions from analysis of talker CL in the LDC emotions data base. (Note: LF parameters expressed as percentage of T0).

	Happy	angry	sad	neutral
Te	46.74	46.43	52.49	53.72
Tp	33.94	33.87	36.71	37.37
Ta	3.61	3.60	3.11	3.21
F0	163.68	176.88	91.04	98.40
AV	56.76	51.97	54.14	54.77

Table 11: Confusion matrix for Linear Discriminant classifier using just the three LF parameters Te, Tp, and Ta to predict emotion.

	happy	angry	neutral	sad
happy	12	9	0	0
angry	7	17	0	2
neutral	0	0	5	12
sad	0	2	6	19

5.5 LF parameters and emotion

In the data obtained from analyses of emotional speech, analysis of variance showed significant main effects for all the LF parameters. Tukey HSD post hoc analysis indicates that for F0, all contrasts except the difference between neutral and sad are significant at $p<0.01$. For Te, there are significant differences between neutral and both happy and anger, and between sad and both happy and anger, but happy and anger are not different, nor are neutral and sad. For Tp, the only significant differences are between sad and happy and between sad and angry. For Ta, the only significant difference is between neutral and angry, although the difference between sad and happy is marginal ($p=.056$). AV is significantly different for happy versus angry and for sad versus angry. BTW, for the HSD contrasts, I am using $p=0.05$ as the cutoff for "significant." The average values as well as the values of F0 are shown in Table 4.

To see what these values mean in terms of the source waveform, Figure 2 shows the glottal flow derivatives generated using parameters associated with each emotion.

Linear discriminant analyses (LDA) was used to see if we could classify emotions based only the source features. Using only the Tp, Te, and Ta parameters, we get about 58% correct classification using leave-one-out cross validation (Table 5). The classification accuracy improves to 75% correct if AV and F0 are also included in the LDA fit (Table 6).

5.6 Discussion

A parameterization of speech based on formants and explicit phonation source features would seem to provide an ideal basis for obtaining features for statistical parametric synthesis, particularly when control of expressive speech properties is an objective. A significant body of research developed over many decades has established formant frequencies as the primary acous-

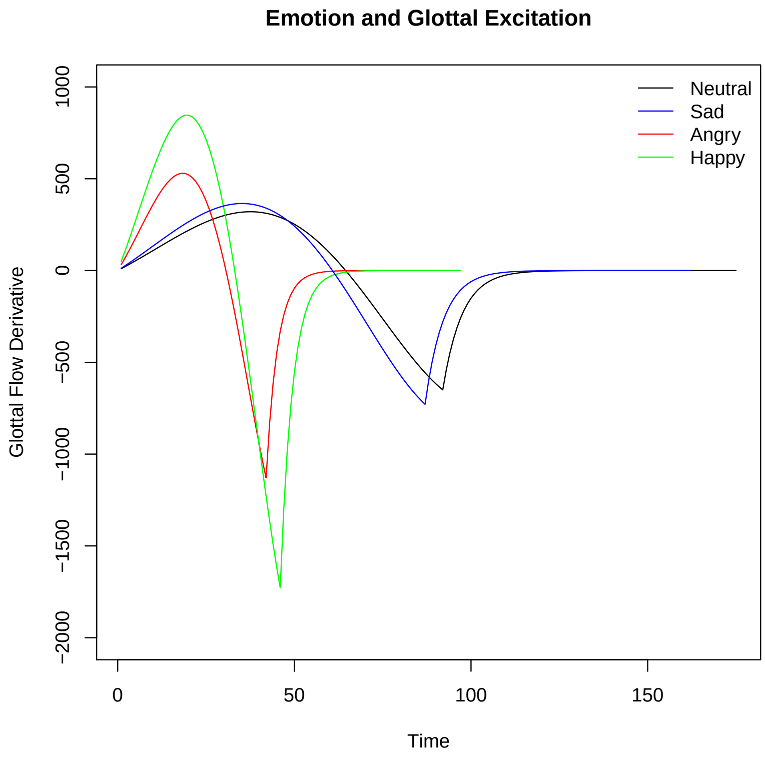


Figure 4: Glottal Flow Derivatives generated by LF model associated with each of four emotions.

Table 12: Confusion matrix for Linear Discriminant classifier using the three LF parameters Te, Tp, and Ta plus F0 and AV to predict emotion.

	happy	angry	neutral	sad
happy	15	4	1	1
angry	4	22	0	2
neutral	0	0	8	9
sad	0	0	4	23

tic features for voiced phonetic segments. Similarly, a substantial body of more recent work has demonstrated clearly that source features like those captured by the LF source model are crucial in representing expressiveness and emotion in speech.

To realize the potential of a formant plus LF parameter representation of speech for statistical parametric synthesis, two things are necessary. First it must be possible to extract the formant and LF parameters from natural speech samples accurately and automatically. It should go without saying that accuracy is essential for any parameterization of speech. For a statistical approach to be practical, it is necessary to derive parameters from a large number of speech samples, which implies that the process must run automatically, or at least with minimal manual input.

The second broad need is for parameters that are amenable to statistical model building. The most successful parameterizations of speech for building statistical models (for synthesis or recognition) such as Mel Cepstral coefficients have the property that they are orthogonal. Orthogonality is not an essential property of parameters, but does allow use of diagonal covariance matrices, which in turn allows stable statistical models to be trained from relatively small numbers of samples. For non-orthogonal parameters, a full covariance matrix is theoretically necessary, requiring estimation of N^2 rather than just N parameters in the model. The formant and LF parameters derived from natural speech are not orthogonal and this suggests that it may be necessary to have a much larger number of training cases to achieve models of similar stability to those obtained from, say Mel Cepstral coefficients.

For the current development effort, ClusterGen models did not produce speech from formant+LF parameters that was comparable to the speech generated from MCP and impulse source excitation. We must assume that this is due to failings in both of the areas identified above. The accuracy of the parameter extraction process was generally improved over the course of the workshop, but still needs improvement. Close observation of the fitting procedure suggests a number of avenues to pursue in future work to improve the fitting procedure. First, the L-M method used in the current algorithm requires calculation of the derivatives of the spectrum with respect to all parameters being estimated. Because it was not possible to analytically estimate derivatives for all parameters, some were estimated using a finite difference approach. The choice of how large a delta to use for the finite differences may have been suboptimal. Use of a different minimization approach (e.g., [28]) could lead to improved fits.

A second area where improvement may be possible with the analysis is how parameter values are constrained to stay within realistic bounds. One problem that was observed in some cases was that the minimization process would drive one or more parameters to physically unrealistic values. Our attempt to use quasi-articulatory (QA) parameters instead of raw formants was one attempt to avoid this problem. The QA parameters were designed to keep parameters within reasonable limits. However, that approach generally led to somewhat poorer copy synthesis than observed for fitting raw parameters. This could be because the parameter constraints were too severe, or possible the unusual parameter set we devised for this trail was suboptimal. These are questions that should be examined further.

Finally, regarding how the formant-LF parameters performed in the context of statistical parametric synthesis, as noted, these parameters are not orthogonal and thus theoretically require use of a full covariance matrix for training the HMM models. However, given the amount of data available, that was not attempted. In future analyses, this problem could be avoided by recoding the parametric data into an orthogonal set of parameters, e.g., by using principal components analysis of the data and training on parameters derived by projecting the raw formant and LF parameters into the PCA space.

Perhaps such changes to the HMM construction process along with additional improvements in the acoustic analysis process we will be able to report improved performance in the future.

6 Modelling 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 [29]. Modeling speech using multiple parallel feature streams allows going beyond the "beads-on-a-string" model [30] of speech, and earlier work shows that AFs are well suited for modeling changes in hyper-articulated speech [31], which we regard as a prototype of a strong emotion. Lisping was also found to clearly affect isolated AFs in speaker adaptation [8].

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.

6.1 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 [29]
- Multi-stream classification: used for recognition – ca. 8 multi-valued individual classifiers [32]
- 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 5 shows the continuous output of the "place of articulation" node of a neural network trained using QuickNet's[4] "continuous" mode using a 0.4 sec. window of stacked MFCC features as input.

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

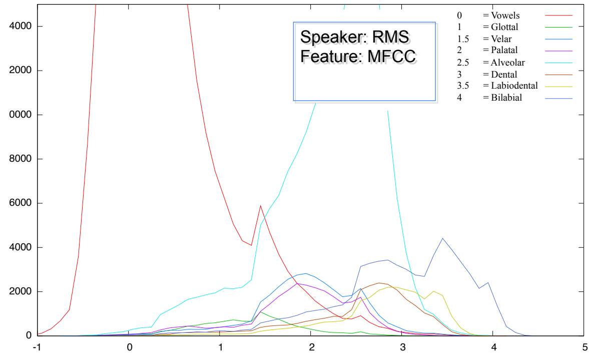


Figure 5: Output distribution (quasi-histogram) of the "front-back" node of an ANN for sounds belonging to different AF categories, trained with the target values shown in the legend. The learned distributions for the 8 classes exhibit inversions of articulatory targets and bi-modal distributions, which, according to manual analysis, mostly stem from improperly labeled, or insufficiently prepared data.

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 5, for comparisons across databases, languages, speaking styles, etc. Similar results have been achieved for other speakers and input representations.

information.

6.2 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 [27] statistical parametric 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 [33]. 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 [34] 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.

6.3 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.

7 Voice Conversion

As a further investigation of using new parameterizations of speech we also applied what we learned during the workshop to the related field of voice conversion.

Voice conversion is the process of modifying speech so that some of the characteristics of the speaker such as age, gender and other aspects of the speaker's identity are changed. Historically, this process has always involved training a model using a small parallel corpus i.e. a small set of utterances (about 30) that are spoken by both the source speaker and the target speaker. A model such as a Gaussian Mixture Model (GMM) [35] is used to learn the characteristics of this source and target speech. This model can then be used to transform the characteristics of the source speaker so that the speech sounds closer to the speech of the target speaker.

However, it is unreasonable to expect that a parallel corpus will always be available especially if the target speaker's speech is hard to collect or if the target speaker is deceased. Therefore a model that can be trained with non-parallel data is be very useful. Another major goal in voice conversion research is to find a representation of speech that perfectly separates the speaker specific characteristics from the linguistic content present. As part of our exploration of new parameters for speech synthesis, we also investigated the possibility that these features might possess some of these properties.

It is well known that speech can be modelled sufficiently by using a Source-Filter model where the filter is derived from a spectral model and the source from an excitation model or pulse & noise. The parameters that we had experimented with at the Johns Hopkins Workshop were the use of Articulatory Features to represent the filter and the Liljencrants-Fant model to represent the excitation. Since the relationship between excitation and speaker identity is unclear, we focused our efforts only on experimenting with the Articulatory Features.

Synthesizing speech directly from articulatory features is non trivial. Therefore it is necessary that we transform the articulatory features into a form more conducive to synthesis. We chose to train a mapping between these articulatory features and Mel Frequency Cepstral Coefficients because the space of MFCCs in synthesis is quite well understood and there are reasonable objective tests that can be used to test quality. We used the Quicknet toolkit to train neural networks for each speaker that would learn a mapping between the AFs and MFCCs for that particular speaker. This mapping is the key to doing voice transformation without parallel data as will be explained in more detail later.

For our preliminary experiments in voice transformation, we wanted to investigate the transform domain i.e. the space in which we perform the modelling of source and target speakers' characteristics. Traditionally, this process uses MFCCs and works by extracting those features for source and target speakers from parallel data; a Gaussian Mixture Model is then built for the joint probability density function of the source and target speakers' MFCCs. This joint model is then used to predict the target speaker's MFCCs. Using the Articulatory Features, We did experiments where we built joint models for the source and target speaker's Articulatory Features and compared it to similar experiments on MFCC joint models.

We used Mel Cepstral Distortion (MCD) as an evaluation metric to measure the quality of voice conversion. We did this by extracting MFCCs from the converted speech of the source speaker and the speech of the target speaker and computing the Euclidean distance between the two after using Dynamic Time Warping to align the two sets of MFCCs in time. The results of these experiments are shown below.

RMS is a male US English speaker, while SLT is a female US English speaker (both databases are part of the CMU Arctic datasets [26]).

GMM on rms MCEP to slt MCEP: 6.35
GMM on rms AF to slt AF: 9.25

In order to be able to do voice transformation with non-parallel data, we made use of the fact that the mapping between articulatory features and MFCCs for a given speaker can be learned without the need for data from any other speaker. Therefore, by training a mapping between the MFCCs and AFs for the source speaker and the reverse mapping (AFs to MFCCs) for the target speaker, it becomes possible for us to do voice transformation with no parallel data. All we need to do for a test utterance is extract the AFs from the source speaker's speech using the source mapper and then put it through the AF-MFCC mapper of the target. We can then use these MFCCs to synthesize speech that sounds like the target speaker. The MCD results for this process is shown below:

MCD using slt AFs on the rms AF-MCEP mapper: 8.68

Although these results are poorer that MCEP to MCEP transformation, we are producing this without parallel data. There we see this technique having great potential, but still clearl deserves further work.

8 Results and Conclusions

The initial goal of the workshop was to investigate alternative parameterizations of speech for statistical parameter speech synthesis. That principal goal has been achieved and the results of using articulatory features and the LF-model show that SPSS modeling techniques can deal with parameterizations beyond classical simple spectral models.

The secondary goal was to move beyond modeling of simple clean read speech. Using expressive speech, both emotion and personality, as the target speech requires both more complex models, and new development of evaluation techniques. The work on using emotion-ID technologies for evaluating synthetic expressive speech is certainly promising, but further work will still need to be done to better understand it correlation with human perception of expressive speech. Though the problem of evaluation of expressive speech is a much larger problem than can be solved in a 6 week workshop.

There were three basic thrusts and the summary of results can be summed up as

AF for speech synthesis extraction, modeling and synthesis

LF-Model extraction, modeling and synthesis

Emotional Synthesis modelling and evaluation framework

But the results are far from full conclusions, we have shown that AFs can used efficiently enough to produce similar quality to MCEPs along in statistical parametric synthesis. Our goals are to investigate this direction further. AFs offer interesting potentials for cross-lingual and inter-speaker conversion. As AF conversion can be run on data without a phoneme set, we have a novel potential direction for modeling of languages without a pre-defined phoneme set or even a well-defined writing syste,.

Articulatory features also have a better notion of continuity than conventional spectral models offering a chance to build "language models" to constrain prediction. That is build a priori models of how AFs change over time.

We have only begun to address the issues of synthesizing expressive speech. The LF-Model shows great potential yet it was only toward the end of the workshop that our prediction system came together. But the success of the small experiments we did do point to an area worth much more investigation.

References

[1] J. Allen, S. Hunnicut, and D. Klatt, *Text-to-speech: The MITalk system*, Cambridge University Press, Cambridge, UK., 1987.

[2] J. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech: A Dynamic Approach*, Springer Verlag, 1993.

[3] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR – ν-TALK speech synthesis system," in *Proceedings of ICSLP 92*, 1992, vol. 1, pp. 483–486.

[4] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP-96*, Atlanta, Georgia, 1996, vol. 1, pp. 373–376.

[5] Heiga. Zen, Keiichi. Tokuda, and Alan. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1059–1064, 2009.

[6] A. Black and K. Tokuda, "The Blizzard Challenge 2005," `http://festvox.org/blizzard/`, 2005.

[7] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the blizzard challenge 2007 listening test results," in *ICASSP 2007*, Bonn, Germany, 2007.

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

[9] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow.," Tech. Rep., STL-QPSR, Speech Music and Heading, Royal Institute of Technology, Stockholm, 1985.

[10] A. Wrench, "The MOCHA-TIMIT articulatory database," Queen Margaret University College, `http://www.cstr.ed.ac.uk/artic/mocha.html`, 1999.

[11] F. Metze, *Articulatory Features for Conversational Speech Recognition*, Ph.D. thesis, Universitaät Karlsruhe, 2005.

[12] S. Steidl, T. Polzehl, T. Bunnell, Y. Dou, P. Muthukumar, D. Perry, K. Prahallad, C. Vaughn, A. Black, and F. Metze, "Emotion identification for evaluation of synthesized emotional speech," in *Proc. Speech Prosody*, 2012.

[13] Florian Eyben, Martin Wöllmer, and Björn 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.

[14] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo 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.

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

[16] D. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müler, and S Narayanan, "The INTERPSEECH 2010 paralinguistic challenge," in *Interspeech*, 2010.

[17] A. Black, T. Bunnell, Y. Dou, P. Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, Steidl S., and C. Vaughn, "Articulatory features for expressive speech synthesis," in *Proc. ICASSP*, 2012.

[18] K. Takeda, K. Abe, and Y. Sagisaka, "On the basic scheme and algorithms in nonuniform unit speech synthesis," in *Talking Machines: Theories, Models, and Designs*, pp. 93–105. Elsevier, Amsterdam, 1992.

[19] Mark Beutnagel, Alistair Conkie, and Ann K. Syrdal, "Diphone synthesis using unit selection," in *SSW3-1998*, Jenolan Caves, Australia, 1998, pp. 185–190, ISCA.

[20] J. N. Holmes, I. G. Mattingly, and J. N. Shearme, "Speech synthesis by rule," *Language and Speech*, vol. 7, no. 3, pp. 127–143, 1964.

[21] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *The Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[22] C Gobl and A Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun*, vol. 40, pp. 189–202, 2003.

[23] H. T. Bunnell and S. P. Eberhardt, "Revisiting analysis by synthesis," *The Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1353, 1999.

[24] J Mahshie and C Gobl, "Effects of varying lf parameters on klsyn88 synthesis.," *Proceedings of the XIVth International Congress of Phonetic Sciences*, pp. 1009–1012, 1999.

[25] C Gobl, E Bennett, and A Chasaide, "Expressive synthesis: How crucial is voice quality?," *Proceedings ICAASP 2002*, pp. 91–94, 2002.

[26] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. CMU-LTI-03-177 `http://festvox.org/cmu_arctic/`, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[27] A. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Interspeech 2006*, Pittsburgh, PA., 2006.

[28] M Frohlich, D Michaelis, and HW. Strube, "Sim–simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals.," *J Acoust Soc Am*, vol. 110, no. 1, pp. 479–488, 2001.

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

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

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

[32] 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.

[33] J. Kominek, T. Schultz, and A. Black, "Synthesizer voice quality on new languages calibrated with mel-cepstral distorion," in *SLTU 2008*, Hanoi, Viet Nam, 2008.

[34] A. Black and J. Kominek, "Optimzing segment label boundaries for statistical speech synthesis," in *ICASSP 2009*, Taipei, Taiwan, 2009.

[35] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," in *Proc. EUROSPEECH95*, Madrid, Spain, 1995, pp. 447–450.