Symbolic vs. acoustics-based style control for expressive unit selection

Ingmar Steiner[1,2], *Marc Schröder*[1], *Marcela Charfuelan*[1], *Annette Klepp*[1,2]

[1]DFKI GmbH, Saarbrücken, Germany
[2]Department of Computational Linguistics & Phonetics, Saarland University, Germany
firstname.lastname@dfki.de

Abstract

The present paper addresses the issue of flexibility in expressive unit selection speech synthesis by using different style selection techniques. We select units from a mixed-style unit selection database, using either forced style switching, no control, symbolic target cost, or acoustic target cost as a style selection criterion. We assess the effect of selection technique, feature weight and relative weight of target vs. join costs on a set of objective measures for style specificity and smoothness.

Index Terms: expressive speech synthesis, unit selection, style control, voice quality, acoustic target cost

1. Introduction

The synthesis of expressive speech is by no means a solved problem [1]. Early approaches used explicit control models to impose emotion-specific prosody in formant or diphone synthesis [2]. The results were recognizable in terms of intended style, but the quality was too unnatural for widespread use. When data-driven synthesis approaches emerged, explicit control was sacrificed in favor of quality. Unit selection synthesis is able to synthesize expressive speech by simply using units recorded while speaking in the intended style [3, 4, 5, 6, 7] – a method that we refer to as a "playback" approach, since no modeling whatsoever is required from the synthesis engine to produce expressive speech. The expressivity in the synthesized output is merely a side-effect of the nature of the recorded material, which in turn means that in order to synthesize a new style, a new database must be recorded. Statistical parametric approaches to expressive synthesis, when they derive their expressivity exclusively from style-specific training [8] or adaptation [9] data, use the "playback" approach as well, in that the expressivity is fully and automatically determined by the data.

The quality of "playback" expressive synthesis can be very good, notably with limited domain unit selection synthesis [4, 7]. A problem of a quantitative nature arises for general domain unit selection: for every speaking style, a very large expressive speech corpus would be needed to produce general text-to-speech with high quality in that style. If it were possible to somehow leverage a large *neutral* speech corpus to improve expressive synthesis quality, the problem would become much less severe. In HMM-based synthesis, this can be achieved using model adaptation techniques [9]; in unit selection, however, there is no simple equivalent of that approach.

However, what is missing in "playback" approaches, both unit selection and HMM-based, is flexibility and control. By flexibility we mean the ability to generate a certain expression only to a certain extent, such as an emotion being conveyed with varying intensity. Control means that at synthesis time, it is possible to trigger a given speaking style, e.g. through speech synthesis markup.

The present paper investigates possibilities for increasing control in expressive unit selection speech synthesis, while limiting the need for large expressive databases.

Previous work on control of expressive unit selection synthesis mainly focuses either on *selection* of suitable units from a mixed-style database, or the *modification* of synthesized speech using signal modification techniques. With respect to expressive unit selection, [10] applied different acoustic selection criteria to retrieve appropriate units from a mixed-style database. In their study, hand-crafted prosodic selection rules outperformed automatically trained HMM-based predictors; the resulting synthesis was perceived as intended for *anger* and *sadness*, but not for *joy*. [11] and [12] used manual annotations of emphasis to select emphatically spoken units when generating emphatically accented syllables in speech. [12] extended this approach to acoustics-based selection by training acoustic models of emphasis on the labeled part of the corpus, and using these models to select candidates from the larger, unlabeled part of the corpus. Results point in the right direction but are not yet fully satisfactory.

Attempts to improve the control of expressivity in unit selection through signal modification have also been reported, using either prosody manipulation (e.g. [13]) or voice conversion (e.g. [14]). The problem with these approaches is the introduction of distortions arising from the signal manipulation, which degrades the quality compared to the unmodified unit selection synthesis.

In the present paper, we address the specific issue of selecting units from a mixed-style unit selection database, using either forced style switching, no control, symbolic target cost, or acoustic target cost as a style selection criterion. We assess the effects of selection technique, feature weight and relative weight of target vs. join costs on a set of objective measures for style specificity and smoothness.

The paper is organized as follows. We first describe the PAVOQUE corpus of German expressive speech (Section 2) that we have used to build expressive voices with different style control methods (Section 3). We then describe our approach to a comparative evaluation of these voices, and to measuring the effects of tuning the cost weights (Section 4), and present the results of this investigation (Section 5). We then discuss these findings and conclude with an outlook of future work.

2. The PAVOQUE expressive speech synthesis corpus

We designed and recorded a German speech synthesis corpus as the basis for the present and other experiments with expressive speech synthesis. The corpus consists of a relatively large body of neutrally spoken speech material, and four smaller expressive parts, all produced by the same speaker. The expressive styles are *cheerful*, *depressed*, *aggressive* (corresponding roughly to the frequently used terms *happy*, *angry*, and *sad* for emotional speaking styles) as well as a “cool, laid back” speaking style recorded for a virtual poker player [7] (referred to as the *poker* style below).

2.1. Prompts

We designed a prompt list consisting of 3 000 sentences, which were automatically pre-selected from the German Wikipedia by a greedy algorithm aiming for best diphone coverage and prosodic variation [15]. The pre-selected sentences were then reviewed, manually corrected and altered where necessary by two phonetically trained research assistants, both native speakers of German. For the expressive voices we used a small subset of 400 sentences selected out of the larger *neutral* prompt list based exclusively on diphone coverage, disregarding prosodic variation. This joint speech material provides at least a minimal amount of diphone coverage for each expressive voice.

In addition, around 150 prompts typical to the respective speaking style were added to each expressive prompt list; the *cheerful* style includes utterances such as, “I am the eternal optimist,” whereas the *depressed* style includes utterances like, “I don’t think you should be so positive,” and the *poker* voice contains utterances such as “I have a Royal Flush.” The motivation for this design was to have high quality expressive voices within domains suitable for the respective expressive style, but sufficient diphone coverage to allow the synthesis of arbitrary text input with reasonable quality.

2.2. Recording

For the recordings, we employed a male professional opera singer, a native speaker of standard German who was also able to produce foreign-language words and phrases naturally and fluently. His voice was very versatile, and he was able to maintain a given expressive speaking style relatively consistently.

The recordings were carried out in a sound-proof room, using a high-quality room microphone positioned approximately 40 cm from the speaker’s mouth. The prompts were presented to him on a computer screen, one at a time, using our recording tool Redstart [16]. This tool can be used to record multiple takes of each prompt, and detects temporal and amplitude clipping automatically. Recordings were made at 24 bit per sample and 44.1 kHz, and later downsampled to 16 bit, 16 kHz. The entire recording procedure was supervised by phonetically trained staff, who requested a repetition whenever necessary.

The speaker was instructed to produce the *neutral* part of the corpus “in a news-reading style”. The expressive speaking styles were defined in terms of the characters from the Sensitive Artificial Listener scenario [17]. The character “Poppy” was described as “nice, optimistic, happy-go-lucky”, “Spike” as “aggressive, irritable and short-tempered” and “Obadiah” as “a wet blanket kind of person”. The speaker was instructed to produce the expressive prompts in-character for each of these personas. No instructions as to how to achieve that effect were given, and he was free to choose voice quality, speaking tempo, intonation, etc. as he saw fit. Before the actual recordings began, the speaker experimented with possible renditions of each style using a number of domain-specific sentences until he and the experimenters felt comfortable with the speaking style used.

Recordings were carried out in sessions of up to five hours, over a period of several weeks. Breaks were taken either at the speaker’s request, or when the supervisors felt that the quality of the recordings was declining (noticeable e.g. by the number of takes required for a sentence, or the speaker “slipping out of character” during the recording of expressive speech).

2.3. Labeling

The recorded utterances were automatically labeled using the German MARY text-to-speech (TTS) phonemizer and forced-aligned with EHMM[1] through the MARY voice building toolkit [16]. The entire corpus was then manually corrected by three phonetically trained research assistants. Both temporal alignment and segmental deviations between the predicted and spoken segment chain were corrected.

The labeling of the expressive material was carried out in the same way as for the *neutral* material. There were no significant differences, but it is interesting to note that the automatic labeling did not perform as well for some expressive material due to voice quality issues (e.g. tense voice in the *aggressive* style and creak in the *poker* style).

3. Building expressive voices

From the recordings described in the previous section, we created a range of unit selection voices differing in the methods used for style control. All voices were built using the MARY voice building toolkit [16] to work with our open source MARY TTS platform.[2]

3.1. Baseline voices

As a baseline for style control, five separate unit selection voices were built: one voice containing speech material from all available styles but without any information about the style, and one for each of the four expressive speaking styles.

Each of the expressive voices used only the 400 phonetically balanced prompts, *without* the additional domain-specific prompts, to ensure comparable unit databases.[3] Since each of these voices was built only from recordings corresponding to the respective speaking style, at synthesis time, style can be strictly enforced by selecting the appropriate voice – we therefore call these voices *forced-style* voices. On the other hand, since each voice was built from less than 400 utterances, smoothness was expected to be fairly low.

As a baseline for smoothness, one voice was built from the neutrally spoken utterances and the four expressive parts of the corpus used to build the *forced-style* voices. This *allstyles* voice uses the full set of 3 014 phonetically balanced *neutral* prompts[3] as well as the four sets of 400 expressive prompts. As with the *forced-style* voices, no information about the style with which a prompt was spoken is available in this voice. Due to the predominance of *neutral* material, this voice was expected to have a higher smoothness than any of the *forced-style* voices; however, in the absence of any style information, it is a matter of chance which style(s) will be used to synthesize a given target utterance.

In these and all other voices, acoustic target weights for duration and F_0, which are normally used in MARY unit selection voices, were set to zero to avoid any confounding effect. Thus, only symbolic target costs were used in the baseline voices.

[1] `http://festvox.org/`

[2] `http://mary.dfki.de/`

[3] Due to labeling mismatches, a small number of utterances had to be removed from this data, and the numbers of utterances from each style actually used for voice building are as follows: *aggressive* 394; *cheerful*, *depressed*, and *poker* 393 each; *neutral* 2 946.

These voices represent the expected extremes regarding style specificity and smoothness, respectively. The interesting question is how style control can maintain a high level of style specificity while improving smoothness.

3.2. Symbolic style target cost

As a first type of style control, we built a voice from the same speech material as the *allstyles* voice, but providing *style* as a target feature. At build time, all units receive a discrete value for this feature, corresponding to the speaking style of the respective source utterance. At synthesis time, speaking style can be controlled by explicitly selecting it for some or all of the input text. This approach, while more flexible than the baseline, nevertheless relies entirely on the symbolic style labels assigned to the individual recordings and is oblivious to the actual acoustic data they contain.

Another drawback lies in the binary target cost: any similarities across speaking styles are ignored, and all styles that differ from the intended style incur the same cost.

We call this voice *symbolic*. The expectation is that, depending on the weight of the style feature relative to the other target costs, and depending on the relative weight of target vs. join costs, this voice's performance will be somewhere between those of the *forced-style* and the *allstyles* voices: when the feature weight is high enough to dominate all other selection criteria, the voice should be similar to the *forced-style* voice of the intended style; conversely, when the feature weight is very low, or when the join cost dominates the target cost, the voice should resemble the *allstyles* voice. The interesting area is at intermediate target feature weights, where the effect of allowing some units that are not in the intended style may be beneficial to smoothness without being too detrimental for style specificity.

3.3. Acoustics-based style target cost

As a more experimental alternative to symbolic selection of style, we have investigated the use of an acoustic feature extracted from the speech data itself. Pitch and spectral measures such as formant frequencies and bandwidths,[4] and spectrum intensity on different frequency bands were used to calculate *voice quality* parameters as described by [18], specifically open quotient gradient (OQG), glottal opening gradient (GOG), skewness gradient (SKG), rate of closure gradient (ROC), and incompleteness of closure (IC). These gradient measures are rough spectral estimates of traditional voice quality parameters normally calculated in the time domain. The voice quality measures were extracted frame synchronously (frame length 25 ms; frame shift 5 ms) from the voiced frames of the data. These gradient voice quality measures have been successfully applied to the classification of emotions [18].

Principal component analysis (PCA) of these parameters shows that 79.2 % of the variance is explained by the first principal component, with a loading of 0.976 for OQG.[5] In our initial approach, we therefore focus on OQG.

The voiced frames in the speech data were used to train a classification and regression tree (CART),[6] which is used at synthesis time to predict an OQG value for each target unit. It was found that using utterance mean OQG values ranked the *style* feature at the top of the resulting CART, confirming our expectations that style and voice quality should be correlated.

The unit selection process was extended with a continuous-value target cost feature representing the CART-predicted OQG value, and a single voice was built with this parameter as the only acoustic target cost.[7] We refer to this voice as *vq* (for voice quality).

4. Evaluation

We carried out a systematic objective evaluation of the different voices built as described above, with the following rationale. Given the fact that the mixed style databases consist of a large *neutral* part and a small expressive part for each of the expressive styles, there is a natural trade-off between style-specificity and smoothness: utterances that are synthesized from the large *neutral* section of the corpus are more likely to find smoothly fitting units, whereas the selection within any of the small expressive sections has only a very limited set of candidate units available. Conversely, the more *neutral* units are used to synthesize an utterance, the less likely it is that the output sounds specific to an expressive intended style.

This reasoning provides us with two types of objective criteria to evaluate the performance of unit selection with an expressive intended style:

(a) the proportion of units from the intended style as a simple measure of style specificity; and

(b) the mean span length (of consecutive units which are adjacent in a source utterance) as an indication of smoothness.

The expected extrema of these criteria are represented by our baseline voices: the *forced-style* voices necessarily produce speech entirely from the intended style, and should sound the least smooth; the *allstyles* voice built from the full corpus but without the style feature is expected to produce the smoothest output but to use predominantly *neutral* material.

The acoustic similarity of units across intended speaking styles is more difficult to assess objectively. Given the importance of voice quality for the intended style, it seems reasonable to assume that a spectral distance measure between a synthesized utterance and a *gold standard* may be able to capture some of the relevant similarities and differences. For this purpose, we employ the same distance measure as previously used in research on voice conversion with the same speech material [19]. As the gold standard, we use the *full set* of 400 utterances recorded in *each* intended style (while dynamically blacklisting the corresponding utterances, see Section 4.3).

In the following, we compare the symbolic and acoustic style control methods with respect to their performance on the two criteria as the weights of the respective style feature and of the join costs are systematically varied. It is expected that for some non-extreme weights, it may be possible to retain a good deal of style specificity while at the same time improving smoothness beyond that of the *forced-style* voices.

4.1. Style specificity

The style specificity measure is computed as the *percentage* of units in a synthesized utterance which come from source utterances recorded in the intended style.

[4] Pitch and formant extraction performed with Snack (`http://www.speech.kth.se/snack/`)

[5] PCA performed with R (`http://www.r-project.org/`) using the covariance matrix

[6] CART building performed with Edinburgh Speech Tools (`http://www.cstr.ed.ac.uk/projects/speech_tools/`)

[7] The symbolic *style* feature target cost was set to zero.

4.2. Smoothness

Units which are adjacent to one another in the recorded speech data incur zero join cost, and if several of such units are selected in sequence, this consecutive span will sound as smooth as possible. We use the *mean span length* (in units) in a synthesized utterance as a simple measure of smoothness.[8]

4.3. Dynamic utterance blacklisting

During evaluation, each synthesized utterance under scrutiny is compared to a gold standard: the original utterance produced by the speaker. If this utterance were available in the unit selection database, there would be a strong bias to select and concatenate only units from that utterance, to the extreme of recreating a perfect copy of the original recording. Comparing such a synthesis result with the gold standard would be meaningless.

To avoid this problem and force units to be selected from different source utterances in the voice data, it is common practice to withhold a test set of utterances, excluding them from the voice building. Consequently, the result will typically sound less smooth and natural than the gold standard, but *how much* less depends on factors such as the voice data and the unit selection itself.

However, a few drawbacks are introduced by this exclusion process. Firstly, it is possible that by removing more than one of the withheld utterances at a time from the voice data, a synthesized test utterance is prevented from selecting units from *any* of these excluded utterances, not just from the single corresponding one. If the set of withheld utterances is chosen differently, units from otherwise excluded utterances would become available and might be selected. Therefore, withholding a set of several unrelated utterances during voice building may influence the unit selection evaluation itself.

Secondly, only those utterances that were withheld can be tested against the gold standard in a meaningful way. Subsequent further testing is only possible after extending the exclusion set and rebuilding the voice, which is typically a lengthy process, especially for large amounts of speech. Furthermore, as a corollary of the previous point, the extended set may exclude units selected in previous tests, invalidating their results.

For these reasons, the present study sidesteps the issue by introducing the notion of selective *utterance blacklisting*. In this flexible approach, the utterances to be excluded are present in the voice data, but their units are not considered as candidates during unit selection. This is controlled dynamically at synthesis time by providing a list of zero or more utterance codes as the blacklist; every candidate unit is then checked and discarded if its source utterance appears in the blacklist.

4.4. Spectral distance from the gold standard

As a spectral distance measure, we used the root-mean-squared error (RMSE) of Bark-scaled line spectral frequency (LSF) values [19]. RMSE between a synthesized style-specific utterance and the respective gold standard is estimated using:

$$RMSE_i = \sqrt{\frac{1}{P}\sum_{k=0}^{P-1}(g_i(k) - s_{m(i)}(k))^2} \qquad (1)$$

where P is the linear prediction (LP) order, g and s are the mapped Bark-scaled LSF vectors of gold-standard and synthesized utterances, respectively, i is the speech frame index and $m(i)$ is the mapping of speech frame indices using phonetic alignment information. An LP order of 18 was used for 16 kHz recordings. The mean RMSE values were computed excluding initial and trailing silence in the signal.

4.5. Varying weights

For the *symbolic* and *vq* voices, we systematically vary the weights to observe the effect on our objective measures. We start with a clear predominance of target costs, by setting the relative weight of target vs. join costs in the overall cost function to 0.95, giving join costs an extremely low contribution to the total cost. This is a setting which in our experience is suboptimal for perceived quality, and has a negative impact on smoothness in particular; we use it only to ensure that we are able to make the style selection feature the dominant factor in unit selection.

Keeping this join cost weight constant, we then vary the target cost weight of the style feature (symbolic *style* in the *symbolic* voice, and the OQG parameter in the *vq* voice), from low to high values. We expect only a very limited effect of style-based selection for low feature weights, and the maximum achievable effect for high feature weights.

5. Results

5.1. Style specificity

The proportion of units chosen from *neutral* or the intended style, or a different style can be seen in Figure 1 for the *symbolic* voice and in Figure 2 for the *vq* voice, for different style feature weights. It can be seen that the *allstyles* voice, which corresponds to either the *symbolic* or the *vq* voice with a weight of 0 on the respective style feature, selects more than 91 % of its units from the *neutral* style, as expected. For the *symbolic* voice, the units are selected mostly from *neutral* or the intended style; the style feature weight has the expected effect, and at weight 100 already completely dominates the selection, with 97 % of units selected from the intended style. Increasing the weight further does not remove the remaining *neutral* units, probably due to the fact that these represent diphones unavailable in the expressive sub-corpora, so they are force-selected from the *neutral* sub-corpus.

The picture for the *vq* voice is quite different (Figure 2). As expected, the acoustic style feature selects units from different styles than the intended one, for all weight values. However, the OQG feature by itself does not succeed, even at high weights, in selecting the majority of units from the intended style. In fact (not shown in Figure 2), only for the *depressed* intended style is a substantial proportion of 25 % of the units selected from the *depressed* source style. This value is already reached at weight 100, and then stays nearly constant. For the other styles, there is no clear success in selecting units from the intended style. This pattern seems to agree with the distribution of OQG values across the styles in the database (Figure 3), where *depressed* is most clearly different from the other styles. Apparently, this *style* feature is not sufficient to distinguish the other styles.

To test the hypothesis that additional acoustic features might be able to add distinctive power to the style selection, we built a variant of the *vq* voice with a style feature weight of 100, that also included the acoustic features *log* F_0 and *duration*, predicted by CARTs which included symbolic style as a predictive feature in a way similar to the OQG parameter. It can be seen (rightmost column in Figure 2) that this slightly shifts the distribution between intended and other styles, but not the

[8] It is inversely correlated with the ratio of join count to unit count.

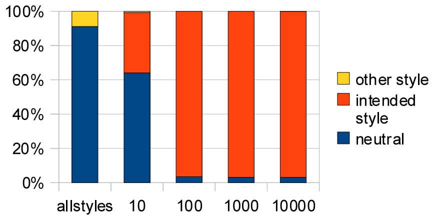

Figure 1: *Unit source style for voice* symbolic *for different style feature weights. Voice* allstyles *corresponds to voice* symbolic *with zero weight on the* style *feature.*

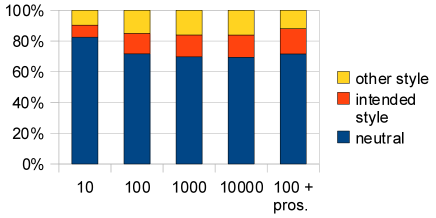

Figure 2: *Unit source style for voice* vq *for different style feature weights. Voice* 100+pros. *corresponds to voice* vq *with weight 100 on the OQG feature and weights for log F_0 and duration tuned such that their contribution to the total target cost is approximately the same as that of OQG.*

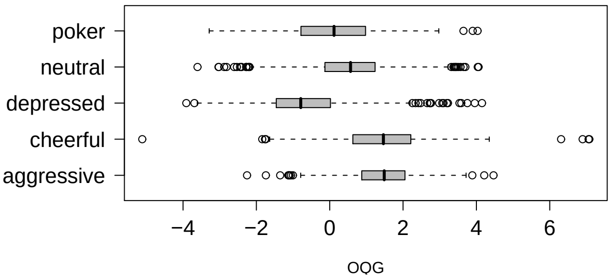

Figure 3: *Distribution of OQG values in the PAVOQUE corpus.*

proportion of *neutral* units chosen. Detailed data per intended style (not shown) reveal that this is due to a substantially higher selection of *aggressive* units as intended compared to the corresponding *vq* voice without prosody features.

5.2. Smoothness

Our measure of smoothness, the mean span length of units adjacent in a single source utterance, is shown in Figure 4. The reference landmarks *allstyles* and *forced-style* are drawn as horizontal lines, which were expected to act as the upper and lower bounds of the smoothness values for the symbolic mixed-style voices.

The expected pattern can be observed for the *symbolic* style for weights 10 and 100. However, mean span length rises again for higher *style* feature weights. We suspect this to be a side effect of the following: all target cost weights are normalized so that they sum to 1. Therefore, a higher weight on one feature lowers the effective weights on all other features. The extremely high weight for the symbolic *style* feature therefore reduces the importance of the other target cost features; given the ceiling effect on the style feature itself, which most of the time produces a cost of 0 (match) for the selected candidate units, this results in an increased importance of the join costs.

The mean span length measure for the *vq* voice exhibits the expected trend, with higher weights on the *style* feature leading to lower mean span length. However, it is unclear at this time why mean span length drops below the baseline for very high weights. Similarly, we currently cannot explain why the *vq+prosody* voice (single point at weight 100) has very long unit spans. More work is needed to investigate this point.

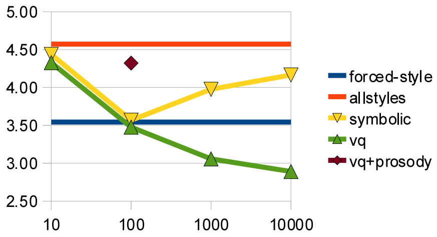

Figure 4: *Mean span length (in halfphone units) for different style feature weights. Constant values for the* forced-style *and* allstyles *voices shown for reference.*

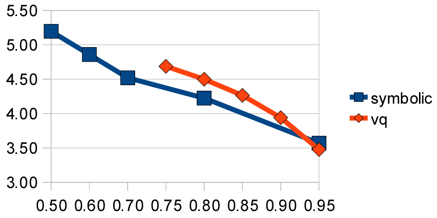

Figure 5: *Effect of changing relative weight of target vs. join costs on mean span length. Style feature weight is kept constant at 100.*

We studied the effect of raising the join cost weight as follows. The *style* target feature weight was kept constant at 100, while the relative weight of target cost vs. join cost was successively reduced from 0.95. The effect on the mean span length can be seen in Figure 5: for both voices, mean span length increases with the join cost weight.

5.3. Spectral distance

The range of expected spectral distances to the gold standard can be seen by comparing the *forced-style* voices (which should be closest to the gold standard) to the *allstyles* voice (which should exhibit the highest distance). Figure 6 shows this pattern for the intended styles *aggressive*, *poker* and *depressed*, but not for *cheerful* utterances in the corpus. Presumably the recorded *cheerful* utterances do not differ spectrally from *neutral* speech in a systematic way.

The effect of the *style* feature weight on spectral distance is shown for *aggressive* speech in Figure 7. For the *symbolic* voice, the distance drops to the lowest expected value from weight 100 and greater; this is expected since most units are

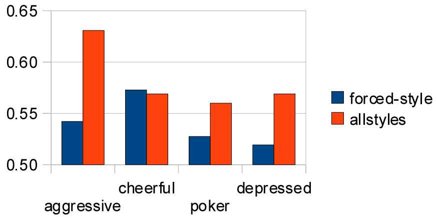

Figure 6: *Spectral distances from gold standard for* forced-style *and* allstyles *voices.*

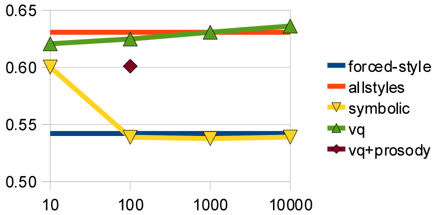

Figure 7: *Spectral distances from gold standard, for* aggressive *intended style and different style feature weights. Constant values for the* forced-style *and* allstyles *voices shown for reference.*

selected from the intended style. However, the distance measure indicates that the *vq* voice does not become more similar to the gold standard with higher *style* feature weights; on the contrary, the distance increases. The variant of the *vq* voice which includes the prosodic features (see Section 5.1), shown as a single data point at weight 100, clearly has a smaller distance than the corresponding version of the *vq* voice without prosody.

6. Discussion and conclusion

We have investigated two different style control techniques: a direct technique, using the intended speaking style as a symbolic target cost feature, and an indirect technique, using acoustic features predicted using a CART. The symbolic *style* feature behaves as expected when its weight is varied. The situation for acoustic *style* features is more complex. The single acoustic feature OQG can partially recover *depressed* units from a mixed-style database, but cannot distinguish the other features; adding F_0 and duration increases the selection of *aggressive* base features. OQG alone does not manage to reduce the spectral distance to the gold standard in the intended style. Adding F_0 and duration predictors seems to help, judging from the single data point that we obtained in this study.

We consider these results to be promising on various levels. On the one hand, the objective measures we have investigated seem to behave meaningfully, even though it is clear that their perceptual relevance remains to be investigated. The use of acoustic style selection criteria appears to deserve further exploration. While the limited feature set we used did not yet prove powerful enough to recover suitable units from the mixed-style database for every intended style, it did select about 25 % of units from the intended style for *aggressive* and *depressed* speaking styles. To what extent this results in a recognizable rendition of these styles remains to be determined. Adding more acoustic features should increase the ability to recover suitable styles. Further work is required to understand the observation that the acoustics-based selection tends to produce smoother synthesis.

7. Acknowledgments

The research leading to these results has received funding from the DFG project PAVOQUE as well as the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231 287 (SSPNet).

8. References

[1] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective Information Processing*, J. Tao and T. Tan, Eds. London: Springer, 2009, pp. 111–126.

[2] ——, "Emotional speech synthesis: A review," in *Proc. Eurospeech*, vol. 1, Aalborg, Denmark, 2001, pp. 561–564.

[3] A. Iida and N. Campbell, "Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders," *Int. J. Speech Tech.*, vol. 6, no. 4, pp. 379–392, 2003.

[4] W. L. Johnson, S. S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," in *Proc. 7th Int. Conf. Spoken Language Processing*, Denver, CO, USA, 2002.

[5] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1099–1108, 2006.

[6] A. W. Black, "Unit selection and emotional speech," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 1649–1652.

[7] P. Gebhard *et al.*, "IDEAS4Games: Building expressive virtual characters for computer games," in *8th Int. Conf. Intelligent Virtual Agents*, Tokyo, Japan, 2008, pp. 426–440.

[8] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 2461–2464.

[9] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 4, Honolulu, HI, USA, 2007, pp. 1233–1236.

[10] N. Campbell and T. Marumoto, "Automatic labelling of voice-quality in speech databases for synthesis," in *Proc. 6th Int. Conf. Spoken Language Processing*, vol. 4, Beijing, China, 2000, pp. 468–471.

[11] V. Strom, R. A. J. Clark, and S. King, "Expressive prosody for unit-selection speech synthesis," in *Proc. 9th Int. Conf. Spoken Language Processing*, Pittsburgh, PA, USA, 2006, pp. 1296–1299.

[12] R. Fernandez and B. Ramabhadran, "Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis," in *Proc. 6th ISCA Speech Synthesis Workshop*, Bonn, Germany, 2007, pp. 34–39.

[13] E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, "Towards emotional speech synthesis: A rule based approach," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, USA, 2004, pp. 219–220.

[14] O. Türk and M. Schröder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 965–973, 2010.

[15] A. Hunecke, "Optimal design of a speech database for unit selection synthesis," Diploma thesis, Saarland University, 2007. http://mary.dfki.de/pavoque/publications/diplomarbeit-annahunecke.pdf

[16] S. Pammi, M. Charfuelan, and M. Schröder, "Multilingual voice creation toolkit for the MARY TTS platform," in *Proc. Int. Conf. Language Resources and Evaluation*, Malta, 2010.

[17] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation," in *2nd Int. Workshop on Emotion: Corpora for Research on Emotion and Affect*, Marrakech, Morocco, 2008, pp. 1–4.

[18] M. Lugger, B. Yang, and W. Wokurek, "Robust estimation of voice quality parameters under real world disturbances," in *Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, France, 2006, pp. 1097–1100.

[19] O. Türk and M. Schröder, "A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 2282–2285.