Exploring Automatic Similarity Measures for Unit Selection Tuning

Daniel Tihelka ⋆, *Jan Romportl* ∗

⋆ Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Republic
∗ SpeechTech, s.r.o., Czech Republic
dtihelka@kky.zcu.cz, jan.romportl@speechtech.cz

Abstract

The present paper focuses on the current handling of target features in the unit selection approach basically requiring huge corpora. In the paper there are outlined possible solutions based on measuring (dis)similarity among prosodic patterns. As the start of research, the feasibility of (dis)similarity estimation is examined on several intuitively chosen measures of acoustic signal which are correlated to perceived similarity obtained from a large-scale listening test.

Index Terms: speech synthesis, unit selection, target features, prosodic patterns, perceived similarity, signal similarity, multidimensional scaling

1. Introduction

Although there is a notable shift towards HMM-based speech synthesis, mainly due its much lower storage requirements and the possibility to express different affective states (not only limited to emotions) or speaker identities, speech generated by unit selection is still generally evaluated as more natural, despite the occurrence of occasional glitches.

One of the biggest problems in unit selection, however, is the coverage of units in all different speaker attitudes or prosodic styles (worse still, affective states). It is estimated in [1] that recording even several hundred thousand sentences will not be enough to guarantee the full coverage of target feature combinations thoroughly describing even basic prosody variability. Therefore, unit selection substitutes the units required to synthesise but not contained in the given speech corpus with units that have the largest match of values in a set of (usually) ad-hoc designed target features and their prominences.

The main problem, we believe, is that "traditional" target cost aims to measure the suprasegmental features (not only prosodic ones in general) of synthesised speech, whereas speech units like diphones cannot, in principle, express any suprasegmental behaviour at all. The target features, thus, assign and fix to the units only those suprasegmental properties which the units had when surrounded by their corpus neighbours in a sequence long enough to express a suprasegmental property. Target cost, as used today, is set to strive for putting the units into the same suprasegmental surroundings as they originally had in the corpus, which is achieved when target features match. This is the cause of the small coverage, resulting in possible glitches when the surroundings exact enough do not exist in the corpus.

As the individual units cannot express any suprasegmental feature, without neighbours each unit may be expected to be able to express a range of different suprasegmental patterns that is larger than the one in which the unit is recorded in the corpus (unit synonymy/homonymy as introduced in [2]). If we were able to arrange each unit in the synthesised phrase to be surrounded by other *appropriate* units, we would get natural sounding speech regardless of the suprasegmental patterns in which each of the units employed originated. How to achieve this is, naturally, neither obvious nor easy, but we suppose that the ability of measuring perceptual similarity of speech is a possible way, as described in the next section.

2. Why to Measure Perceptual Similarity

The idea of the relation between perceptual similarity measure and unit selection is rather simple: let us have a suprasegmental pattern[1] pronounced by a speaker in his/her natural manner, and let us somehow exchange units in the pattern. If there are variants of the original pattern sounding perceptually similar, target properties of units in those new variants can be extended to reflect the new positions into which the units fit (units homonymy). Or, looking at the idea from the other side, the set of all thinkable and feasible target descriptions of each unit in all their homonymous positions (within the natural-sounding variants) can be analysed in order to find a minimal set of the most descriptive and mutually uncorrelated features. For that set of features, the target cost for a unit in any of its homonymous positions should be considerably lower than it is for other positionings of the unit.

One of the problems with similarity perception on speech signal is that speech is of varying nature, passing sequentially through time (contrary to images or various stimuli represented by text, on which similarity perception has already been studied e.g. in [3, 4]). Therefore, it is basically impossible to authentically evaluate or measure how similar two variants of a whole phrase sound. Keeping this in mind, we decided to identify prosodic patterns with *prosodic words*[2] (sometimes also called phonemic words), which are short enough to obtain reliable human evaluation of how similarly any two variants sound, and, simultaneously, they are considered natural constituents of rhythmic and prosodic structure in Czech [5] (and such a kind of prosodic constituents is common to most of rhythm-based languages). Also, we can afford to process each pattern (prosodic word) independently. It may seem that this independence is likely to result in the loss of relation to the overall prosody of the whole phrase, and, as a consequence, prosodic patterns could potentially be placed randomly through the phrase during synthesis (regardless of the fact that each individually sounds natural), which would then lead the synthesised phrase to ex-

This research is supported by grants GACR 102/09/0989 and MSMT 2C06020

[1] For the purposes of this research, suprasegmental prosodic pattern is defined as the sequence of speech units which constitutes the perception of the rhythmic, intonation and phonation qualities of speech; moreover, the pattern does not only consist of prosodic description but it is a part of speech which can be listened, understood and evaluated.

[2] To keep the generality of descriptions, the term *pattern* will still be used in the paper, always referable to prosodic words, though.

press a corrupted communication function, if functionally involved units were placed somewhere in the phrase except the determined position. However, the positioning of patterns within the phrase can be kept as an extra feature, incorporating position diversity into homonymy description in cases when a unit instance is equally used in different patterns within phrases.

Let us outline the future practical utilisation of perceptual similarity measure, supposing a reliable measure at our disposal. Being inspired by [6], a large number of variants of a natural pattern (i.e. prosodic word existing in the corpus) can be built by the raw concatenation of unit sequences with various length, while any part of the natural pattern is excluded from use. A sufficient number of those variants sounding the most similar (and natural, therefore) to the original pattern, as determined by the similarity measure technique, would be considered for further analysis. In it, the features feasible for the target description of each unit in its new contexts (given by the actual position in particular variants) might be examined with the aim to either extend the set of possible target descriptions of a unit, or to include only features relevant for the description of a unit matching all the variants where it appeared.
An alternative is to create the variants by replacing a single unit in the natural pattern instead of generating the whole pattern. Again, the variants sounding the most similar to the pattern would be determined and analysed in the same way as in the first case, with the difference that only one unit will be examined in each variant. Although it may seem to lead to a simpler measure of similarity, due to the variants being equal to the original except for that one unit, the number of patterns to analyse is much higher. Also, the building of the reference similarity perception model is much harder (if at all possible), since many more patterns must be evaluated in listening tests – see Section 3.1. And finally, it is arguable how capable this approach is to model the reality, as the form of unit sequences determined by unit selection is more similar to the way of variants building in the above consideration than to this model. Therefore, the paper is rather focused on employing the whole patterns, while we expect that the experience we gained with similarity measure is fairly general and adaptable for both kinds of approaches, whichever will be chosen.

The approach also poses difficulty regarding the danger of combinatorial explosion. A reasonable, though not optimal, solution is to examine the similarity on a reasonable subset of variants chosen at random (as also proposed and discussed in [6]), and to use massive parallel or grid super computing.

3. The Measure of Perceptual Similarity

To formalise further reading, let A and B be realisations of two prosodic patterns. Let then $\widetilde{s}(A, B)$ be their *measurable similarity* computed on the basis of the *measurable properties* of the patterns (e.g. their signal), and let $s(A, B)$ be their *perceived similarity* representing unmeasurable true reality how similarly A and B are perceived by humans. For practical purposes, it is, however, simpler to work in terms of *dissimilarity* – for measurable dissimilarity it can be defined $\widetilde{d}(A, A) = 0$, whereas $\widetilde{s}(A, A) \rightarrow \mathcal{Z}$ (where $\mathcal{Z}$ can vaguely be defined as a positive number sufficiently large), and it can be considered that $\widetilde{d}(A, B) \approx \mathcal{Z} - \widetilde{s}(A, B)$. In the case of perceived dissimilarity, the situation is not so straightforward, as there is no guarantee of perceived dissimilarity being a symmetric counterpart of the similarity. In [4] the authors showed, using Tversky's contrast model [7], that people tend to attend more to common features of stimuli when evaluating similarity and to distinctive features when evaluating difference. It may cause an object pair to be evaluated as more similar and more different at the same time, if compared to the same evaluation of another pair. However, in the case of acoustic stimuli (evaluated pair of prosodic words in our case) comparison, we presume that the existence of a perceptually distinctive feature in compared patterns is likely to imply higher dissimilarity than it would when similarity is evaluated. Without evidence, human acoustic perception seems to be better in distinguishing difference (it is easier) than in recognising similarity.

Let us now assume that there is a deterministic relation between the two dissimilarities

$$\mathcal{F} : d(A, B) \rightarrow \widetilde{d}(A, B) \qquad \forall A, B \tag{1}$$

thus having a data set with known behaviour of d, we need to find such $\widetilde{d}$ which is *significantly correlated* with the d. Then, we can use $\widetilde{d}$ to estimate d for data not found in the dataset.

3.1. Perceived Dissimilarity

We assumed in Equation (1) that we have $d(A, B)$ at our disposal. However, to be exact, what we only have is its estimate obtained by listening tests[3], averaging the different opinions (judgements) of people regarding what sounds similar and what does not (and to what extent), expecting an "objectiveness" to emerge on the basis of cross-listener agreement.

To obtain the dissimilarity judgements, we carried out listening tests described in detail in [8]. There were 63 listeners participating in them, each evaluating the level of dissimilarity on 780 pairs (including some repeated for validation) combined from 17 prosodic words. Where possible, the words were chosen so that their variants covered different positions in phrases and different melody patterns with at least two examples in each. The signals of the variants were obtained from a female corpus recorded for our TTS system ARTIC [9], each word cut on boundaries given by automatic segmentation, manually checked and faded in and out to suppress the influence of surrounding words. The listening tests were carried out through specially developed web application, and they were financially well-rewarded. Each participant has been familiarised in detail with the purposes of the tests as well as with the examples delimiting exemplary evaluations. The levels of dissimilarity feeling were defined as

- *clearly dissimilar* – clear after the very first listening,
- *dissimilar* – quite close but still recognisably not the same,
- *quite similar/indistinguishable* – being very close even if differing after careful listening, or not recognisable at all.

The dissimilarity was requested to be evaluated for four different aspects (see [8]), while the paper focuses on *overall dissimilarity* intended to evaluate *difference as such, on all the qualitative levels on which the acoustics are perceived and a difference can be felt*. During results inspection we confirmed that for each word at least one pair sounding *quite similar/indistinguishable* exists.

To obtain a (dimensionless) value representing dissimilarity $d(A, B)$, $\forall A, B$, non-metric multidimensional scaling (MDS) of the listening test results was carried out (all variants of one prosodic word analysed at once, although independently for

[3] In [3, 10], the dissimilarity evaluation obtained in the form of a set of listener responses is referred to judged dissimilarity.

each of the prosodic words). This technique, also called *geometric model* [10], assumes that a perceptual effect on evaluated stimuli is inversely related to their distance in a n–dimensional space; it has been used for quite a long time in cognitive science, and for the first time in [11] for synthetic speech quality evaluation. The dissimilarity matrix required by MDS was created in such a way that each cell represented the number of times when a pair of prosodic words was perceived as *clearly dissimilar*, plus the half of the number of times when the pair was perceived as *dissimilar*. The dimension n was chosen ad-hoc to 3, the experiments with other dimensions had no significant impact on the results, though. The dissimilarity estimate $d(A, B)$ was then simply computed as Euclidean distance between stimuli A and B in the 3D space. Although there is some evidence that all the distance axioms valid in metric space are not necessarily always valid in the perception [7, 10], this is, however, not considered in this experiment.

3.2. Measurable Dissimilarity

Let us expect, for the purposes of this paper, that the perceived dissimilarity $d(A, B)$ in Equation (1) matches the dissimilarity really perceived by humans as closely as possible. Now we need to find such a measure on signal which for each pair of prosodic patterns A, B (prosodic words) would return a value $\widetilde{d}(A, B)$ significantly correlated with $d(A, B)$.

In our first attempt, we have chosen pitch-synchronous analysis of compared patterns, with frames of speech signal two pitch-periods long. We can define measurable distance between frames i and j as $\widetilde{d}_{ij}(A, B)$, where $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$ are the number of frames in patterns A and B. To obtain the measurable dissimilarity of the whole patterns, we have chosen symmetric DTW algorithm slightly modified to ensure that only frames from the same phones are compared together, with overlap to $1/3$ of the preceding and following phone allowed. The measurable dissimilarity is thus defined as

$$\widetilde{d}(A, B) = \min_{\{\mathcal{I}(k), \mathcal{J}(k), K\}} \left(\sum_{k}^{K} (\widetilde{d}_{\mathcal{I}(k), \mathcal{J}(k)}(A, B) * w_k) \right) \tag{2}$$

where $\mathcal{I}(k)$ and $\mathcal{J}(k)$ are functions warping k^{th} step in DTW into coordinates of compared frames in the plane spanned by A and B pattens (i.e. $\mathcal{I}(k) = 1, \ldots, I$ for $k = 1, \ldots, K$), Weight $w_{i,j}$ is path penalty encouraging diagonal steps $w_k = 1,\ \mathcal{I}(k) \neq \mathcal{I}(k-1) \wedge \mathcal{J}(k) \neq \mathcal{J}(k-1)$, and penalising steps up and right $w_k = 2,\ \mathcal{I}(k) = \mathcal{I}(k-1) \vee \mathcal{J}(k) = \mathcal{J}(k-1)$. As there is requirement for $\widetilde{d}(A, A) = 0$, it must be true that $\widetilde{d}_{ij}(A, A) = 0, \forall i = j$.

In the present paper we have experimented with the following, intuitively chosen, methods for $\widetilde{d}_{ij}$ computing (note that the compared patterns are always different variants of one prosodic word):

Waveform dissimilarity (WFD) was motivated by the presumption that signal (dis)similarity is likely to imply perceived (dis)similarity, at least for voiced phones. Thus, to measure the dissimilarity on voiced frames, cross-correlation was applied on i, j pairs of von Hann window-weighted frames. For unvoiced segments, where it is meaningless to correlate signals, the dissimilarity was estimated simply by the ratio of the zero-crossing values of frames + their ratio of RMS, which led to values also in $\langle 0, 1\rangle$ interval. The comparison of voiced and unvoiced segments resulted in maximum dissimilarity value 1.0.

Spectra comparison was chosen as an alternative to cross-correlation, as it may be objected that the comparison of the spectral properties of signal is more likely to aptly capture similarity of the signals. In the paper we present similarity computed as the Euclidean distance between the raw magnitude of FFT points of frames (FFT$_r$). Besides, we have also attempted to weigh the spectrum by equal loudness contour (FFT$_l$) to follow how magnitudes of different frequencies affect human hearing, and to smooth the spectra by moving average filter of 10 points (FFT^s_r and FFT^s_l). However, it either did not lead to a significant difference, or the results were even worse.

Singular value decomposition (SVD) is an alternative approach employed in [12] for the measure of join cost. The author showed that the decomposition can be considered as an alternative to magnitude spectrum, which may not explicitly expose a frequency, but it contains both power and phase information "encoded" in the values. Moreover, compared to the spectrum, it is also localised in time (uses frames) but global in scope, as all frames (from all variants of one prosodic word in our case) are decomposed using the same transform kernel. $\widetilde{d}_{ij}(A, B)$ was computed as the cosine of the angle between transformed frames, as inferred in [12].

MFCCs are widely used also in unit selection technique for the measure of concatenation smoothness, and they were even used for the measure of distance betwen units (of the same phone type) in a well-known work [13]. To get $\widetilde{d}_{ij}(A, B)$, Euclidean distance between 12 coefficient MFCC vectors of the given frames was used (MFCC$_E$), as it often appears in concatenation cost. Similar to [13], the dissimilarity was also computed as Mahalanobis distance (MFCC$_M$), which displayed slightly worse results, however.

4. Results

For each pair of all the 780 pairwise combinations of the variants built from the given 17 prosodic words, the perceived dissimilarity $d(A, B)$ was determined from MDS representation of the corresponding prosodic words. Measurable distance $\widetilde{d}(A, B)$ was also computed using the chosen measures between each of the same pairs, and correlated to $d(A, B)$. In Table 1, the correlations are summarised, together with the number of variants (patterns) being combined – for n variants, the correlation was computed from $n(n-1)/2$ pairs of $d(A, B)$ and $\widetilde{d}(A, B)$ values.

It can be seen that the correlations obtained are ranging from highly correlated to virtually uncorrelated. We expect that it is not related to a character of the stimuli, as there are no significant comparable tendencies in the phonetic structure of the most "successful" prosodic words. It is worth noting that although SVD, as presented in [12], did not distinguish between the voicedness of segments, the use of unvoiced segments measure in the same way as for WFD slightly increased correlation to 0.531. For other types of measure it has not been examined, as they provide distance values far higher than $\langle 0, 1\rangle$ interval.

Let us see how the individual measures are correlated, and which are thus likely to compare similar properties of speech. In Table 2, the correlation of $\widetilde{d}(A, B)$ values for all 780 word-pairs are shown. The values indicate that MFCC and waveform-based measures seem to capture similar properties of the signal. The FFT$^s_\cdot$, MFCC$_M$ or SVD modifications of measures displayed correlation higher than 0.95 and thus they are not included in the table.

To obtain an idea about the validity of results we need to have some knowledge about the reliability of the reference eval-

Table 1: *The correlation of* $\widetilde{d}(A,B)$ *to* $d(A,B)$ *for "the best" measures. The* words *are printed in SAMPA alphabet, each having patterns no. variants being combined together.*

Patterns no. $+$ **word**	**WFD**	**FFT**$_r$	**SVD**	**MFCC**$_E$
6 spolupra:t_se	0.658	0.511	0.671	0.597
6 pru:mislovi:x	0.888	0.512	0.666	0.388
6 potravin	0.169	0.847	0.085	0.633
7 za:kazJi:ku:	0.562	0.814	0.516	0.581
7 vminulosci	0.765	0.888	0.640	0.783
7 nasvjece	0.904	0.866	0.689	0.888
8 novina:P\u:m	0.743	0.681	0.715	0.740
8 nemu:Zeme	0.478	0.550	0.444	0.273
9 pozornost	0.773	0.831	0.819	0.831
10 informat_si:	0.757	0.765	0.808	0.653
11 konkurent_se	0.526	0.703	0.284	0.467
11 t_Slovjeka	0.689	0.696	0.579	0.576
12 hospoda:P\stvi:	0.267	0.532	0.071	0.344
13 republit_se	0.344	0.709	0.288	0.445
13 rospot_Stu	0.360	0.742	0.195	0.536
14 proble:mu:	0.674	0.639	0.544	0.736
14 zdu:razJil	0.106	0.492	0.113	0.209
Average:	0.568	0.693	0.478	0.569

Table 2: *The correlation of some of* $\widetilde{d}(A,B)$ *measures.*

	WFD	**FFT**$_r$	**FFT**$_l$	**SVD**
MFCC$_E$	0.907	0.797	0.734	0.818
SVD	0.809	0.599	0.694	
FFT$_l$	0.601	0.855		
FFT$_r$	0.682			

uation, which is a topic discussed in [8] – although the $d(A,B)$ can only emerge from a wider range of subjective opinions, there still may be unaccountable[4] answers in the test distorting the measure. Moreover, even when there is non-accidental cross-participant agreement in evaluations (mentioned in Section 5), it does not necessary imply that listeners evaluated what we intended them to do. And with unreliable $d(A,B)$ we cannot determine if $\widetilde{d}(A,B)$ measures similarity, even if it really does. Therefore, we need to confirm which of the dissimilarities is likely to fail, i.e. if $d(A,B)$ is close to *quite similar* but $\widetilde{d}(A,B)$ is close to *clearly dissimilar*, and we perceive A and B as very similar, it is $d(A,B)$ which is correct. We focused on WFD and FFT$_r$, and analysed five to eight AB pairs with the most different $d(A,B)$ and $\widetilde{d}(A,B)$ values from three the most uncorrelated word pairs, and three such AB pairs from the other pairs – in other words, we checked if $d(A,B)$ is what we hear in cases where $\widetilde{d}(A,B)$ is in disagreement. In 109 pairs in total, with 13 items overlapping for both WFD and FFT, we have found that $d(A,B)$ was correct (or close to our subjective opinion) in 103 cases (94%) for WFD, and in 98 cases (90%) for FFT. It confirms that presented $\widetilde{d}(A,B)$ are not robust measures.

5. Conclusion

It can be seen that, unfortunately, none of the intuitively chosen measures examined provides a basis robust enough to be used for target features definition as proposed in Section 2, the most questionable being the variability among individual words. Other ways of $\widetilde{d}(A,B)$ measure must, therefore, be examined – either they are based on the idea of perceived dissimilarity being a deterministic consequence of signal dissimilarity, or they are inspired by Tversky's feature contrast model [7] or fuzzy feature contrast model [3] (it will, however, require the definition of predicates). Moreover, the behaviour of dissimilarity measures should also be verified on other voice(s).

Although it was shown that $d(A,B)$ is close to our subjective opinion in the majority of examined cases, careful verification of listener responses aiming to determine unaccountable answers is crucial – the cross-participant agreement computed by means of Fleiss' kappa [14] is only 0.21, which may be enough to confirm (on significance level 0.05) that the observed agreement is not accidental, and, realising the vague nature of similarity, it shows that a phenomenon of prosodic patterns similarity is perceived by humans. However, our preliminary experiment, where 18 randomly chosen test participants revised their answers with the highest likelihood of miss (determined as described in [8]), increased the kappa at least by 0.1.

Finally, when building the MDS dissimilarity matrix, there is also the possibility to employ the likelihoods of evaluation confidence for individual listening test participants [8]. The choice of MDS technique should be critically revised as well.

6. References

[1] V. Strom, R. Clark, S. King, S. "Expressive prosody for unit selection speech synthesis", in Proc. of Interspeech, pp. 1296–1299, USA, 2006.

[2] D. Tihelka, J. Matoušek, "Unit selection and its relation to symbolic prosody: a new approach", in Proc. of Interspeech 2006, pp. 2042–2045, 2006.

[3] S. Sanitini, R. Jain "Similarity measures". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 871–883, 1999

[4] A. Tversky, I. Gati, "Studies of similarity", *Cognition and Categorization*, pp. 79–98, 1978.

[5] J. Romportl, J. Matoušek, D. Tihelka, "Advanced prosody modelling", *Lecture Notes in Artificial Intelligence*, vol. 3206, pp. 441–447, 2004

[6] Hélène François, Oliver Boeffard. "Evaluation of Unit Selection Criteria in Corpus-based Speech Synthesis", in Proc. of Eurospeech, pp. 1325–1328, 2003.

[7] A. Tversky, "Features of similarity", *Psychological Review*, 84, pp. 327–352, 1977.

[8] D. Tihelka, J. Romportl, "Statistical Evaluation of Reliability of Large Scale Listening Tests", in Proc. of ICSP 2008, pp. 631–636, 2008.

[9] Matoušek, J., Romportl, J., Tihelka, D., and Tychtl, Z. "Recent Improvements on ARTIC: Czech Text-to-Speech System", in Proc. of ICSLP, vol. III, pp. 1933–1936, 2004.

[10] F.G. Ashby and N.A. Perrin, "Towards a unified theory of similarity and recognition", *Psychological Review*, vol. 95, no. 1, pp 124–50, Jan 1988.

[11] C. Mayo, R. Clark, S. King. "A Multidimensional Scaling of Listener Responses to Synthetic Speech", in Proc. of Interspeech 2005, pp. 1725–1728, 2005.

[12] J.R. Bellegarda, "A novel discontinuity metric for unit selection text-to-speech synthesis", in Proc. SSW5-2004, pp. 133-138, 2004

[13] A.W. Black, P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis", in Proc. of Eurospeech, pp. 601–604, 1997.

[14] J.L. Fleiss, "Measuring nominal scale agreement among many raters", in Psychological Bulletin, vol. 76, no. 5, pp. 378–382, 1971.

[4] The principle of listening tests does not consider any answer as "bad", but there may be answers which are in clear disagreement with test instructions.