Towards Automatic Measure of Similarity for Use in Unit Selection

Daniel Tihelka
Depart. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Republic
dtihelka@kky.zcu.cz

Abstract

The present paper focuses on the unit selection approach to speech synthesis, discussing drawbacks mainly related to the current handling of target features that basically results in the need of huge corpora. In the paper there are outlined possible solutions based on measuring (dis)similarity among prosodic patterns. In the initial experiment, trying to verify the feasibility of the proposed solution, the (dis)similarity of acoustic signal measured by different techniques is correlated to perceived similarity estimate obtained from a large-scale listening test.

1. Introduction

Although the unit selection speech synthesis approach is still in the centre of researchers' attention, there is a notable shift towards HMM-based speech synthesis. This is due to the fact that the HMM approach requires much lower disc space and allows modifying speech either to express different affective states (not only limited to emotions) or even speaker identity. On the other hand, although HMM-approach based speech sounds more smooth, speech generated by unit selection is still generally evaluated as more natural, despite the occurrence of occasional glitches.

One of the biggest problems in the unit selection approach is the coverage of units in all different speaker attitudes and prosodic styles (or even worse, 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. As this is never possible (nor meaningful) to achieve, unit selection substitutes the units not matching target specification with the closest substitutes, where the "closeness" is defined by (usually) ad-hoc designed weights for individual sub-features in target cost – the lower the weight for a feature, the more easily a unit can be replaced by another unit.

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 do not express (they cannot in principle) any suprasegmental behaviour at all. The target features assign (and fix) to the units only those suprasegmental properties which the units had when surrounded by their neighbours in the corpus, creating a sequence long enough to express the suprasegmental property. Target cost as used today is set to strive, together with join cost, for putting the units into the same suprasegmental surrounding as they originally had in the corpus, which is achieved when target features match. This is the cause of the small coverage, resulting in glitches when the surrounding did not exist in the corpus.

As the individual units cannot express any suprasegmental feature, each unit can in general be used to express a larger spectrum of different suprasegmental patterns than the one in which the unit originally existed (unit synonymy/homonymy as introduced in [2]); it is "only" necessary to arrange each unit in synthetic speech to be surrounded by other appropriate units. How to achieve this is, naturally, neither obvious nor easy, but we suppose that the ability of measuring the perceptual similarity of speech can help.

2. Why to Measure Perceptual Similarity

Let us describe here how the ability to measure perceptual similarity can be used to find those suprasegmental patterns[1] in which a unit can be used (not all will obviously be found in practise, as outlined later), while the whole pattern still sounds natural. The idea is rather simple: let us have a set of suprasegmental patterns pro-

[1] For the purposes of this research, suprasegmental prosodic pattern can be defined as the sequence of speech units which constitutes the perception of the rhythmic, intonation and phonation qualities of speech.

nounced by speaker in his/her natural manner, and let us somehow exchange units in the patterns. If the new patterns sound perceptually similar to the their original samples, the target properties of units can be extended to reflect the new possibilities of the units use (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 can be analysed in order to find such a minimal set of features that gives much lower target cost to units in their homonymous positions (not necessarily always the same cost) than it gives to other positions. The aim is to tune the selection algorithm in such a way that a *synthetic version of a phrase in the corpus* will be perceived as similar as possible to the natural recording of that phrase (when no unit from the original phrase is used, modelling a situation whereby the phrase was never recorded). Then we can expect that the synthesised phrases not recorded in the corpus will also sound similar to their representative natural realisations (how a phrase would be pronounced by the speaker if it were pronounced) in the same way as the phrases used in the model situation sounded.

At the moment, we are considering the approach inspired by [3], where the phrases to synthesise are decomposed into all the possible sequences of units of various lengths (forming overlapped parallel sequences leading from first phone to the last), each sequence is "synthesised"[2] using several instances of corresponding units, and then analysed. We also propose to generate every pattern from the corpus in many (all, in the ideal case) "synthetic" realisations which are built by combining a large number (all, ideally) of instances of a large number (a sufficiently large number of those the least overlapping) of decompositions. The consecutive analysis of all those realisations which sound *similar* to their natural paradigms (which implies that they all sound natural as well) will be aimed to find the features pertaining to the homonymy relation for those unit instances which appear on various positions in the realisations. Naturally, each unit instance must be used to create a number of the realisations.

There is also an alternative possibility to generate the "synthetic" realisations by replacing only one unit in each natural pattern (instead of generating the whole pattern), while the analysis of the patterns sounding most similar to the natural ones can then be carried out with the same aim as already described. Although it may seem to lead to a simpler measure of similarity due to the synthetic realisations being equal to the original except for that one unit, we have not yet thoroughly analysed the pros and cons of the individual approaches. Therefore, the paper is rather focused on employing 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.

One of the problems with similarity perception on speech signal is that it can only be meaningfully observed and measured on acoustic stimuli shorter in duration – contrary to images, speech is of varying nature, passing sequentially through time (it is thus rather impossible to authentically evaluate how similar two variants of a whole phrase sound). Therefore, we decided to consider prosodic patterns equal to *prosodic words*[3] (sometimes also called phonemic words), which are considered the natural constituents of rhythmic and prosodic structure in Czech [4] – something similar determining such structures is likely to exist in many other languages. Contrary to [3], each pattern (prosodic word) can then be processed independently, which significantly reduces the number of decompositions examined. It may seem that independence in processing is likely to lose the relation to the overall prosody of the whole phrase, and, as a potential consequence, a basically random placement of prosodic patterns (regardless of the fact that each individually sounds natural) through the phrase during synthesis will cause conflicts between the semantic content (given by phones) and the communication function (given by overall phrase intonation). However, the position of patterns within the phrase can be kept, incorporating position diversity into homonymy description in cases when a unit instance is equally used in different patterns within phrases.

Another problem with the approaches is the danger of combinatorial explosion. A reasonable, though not optimal, solution is to examine the similarity to the original pattern on a reasonable subset of possible "synthetic" realisations of the pattern, chosen at random (as proposed and discussed in [3]), and also 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 (the signals of prosodic words). 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

[2] meaning just concatenating unit sequences to create a given decomposition

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

$\widetilde{s}(A,A) \rightarrow \mathcal{Z}$ (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 [5] the authors showed, using Tversky's contrast model [6], 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 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 *highly correlated* with the d. Then, we can use $\widetilde{d}$ to estimate d for data not found in the dataset. Simply said, this is the aim of our research.

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[4], averaging the different opinions (judgements) of people regarding what sounds similar and what does not (and to what extent), expecting that an "objectiveness" emerges on the basis of cross-listener agreements.

To obtain the dissimilarity judgements, we carried out listening tests described in detail in [9]. There were 63 listeners participating in them, each evaluating the level of dissimilarity on 780 pairs combined from 17 prosodic words – if 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 prosodic words were obtained from a female corpus recorded for our TTS system ARTIC [10], 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 due to quite a large size of the test, the participation (and correct finishing) was 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,

and the dissimilarity was requested to be evaluated for all of the following aspects (resulting in 4 values)

- *timing* – difference in shortening or lengthening rhythm through the prosodic words,
- *intonation* – difference in melody course or slope through prosodic words (not overall pitch level, though),
- *voice colour + pitch level* – difference in voice colour and/or overall pitch level as such,
- *overall feeling* – difference as such, on all the qualitative levels on which the acoustics is perceived and a difference can be felt.

The categories were chosen on the basis of the listening tests data analysis before the tests were started, as well as of our intuitive reasoning, and they are aimed for the study of the dissimilarity relations/prominences of distinctive prosodic constituents (e.g. by Tversky's feature contrast model [6]). Moreover, the variants of a prosodic word in all test stimuli were presented in the order AB to one half of the listeners, and in order BA to the other half (selected at random), to study the occurrences of dissimilarity asymmetry [5, 7]. Nevertheless, neither all evaluated aspects except the overall dissimilarity, nor the asymmetry have been analysed yet.

To obtain a (dimensionless) value representing dissimilarity $d(A,B),\ \forall A,B$, non-metric multidimensional scaling (MDS) of the listening tests results was carried out (all variants of one prosodic word analysed at once, although independently for each of prosodic words). This technique has been used for quite a long time in cognitive science (so called *geometric model* [7], assuming that a perceptual effect on stimuli is inversely related to their distance in a n–dimensional space), but for the first time it was used for synthetic speech quality evaluation in [11]. MDS allows representing the stimuli used for listeners' dissimilarity judgements (individual variants of a prosodic word) as points in n–dimensional space which is configured so that more similarly perceived stimuli pairs are placed closer together, while less

[4] In [7, 8], the dissimilarity evaluation obtained in the form of a set of listener responses is referred to judged dissimilarity.

similar are placed further apart. 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 adhoc to 3 for us to be able to visually analyse and interpret stimuli distribution (not presented in this paper, though); the optimality of dimension choice was not checked at the time of writing. The dissimilarity estimate $d(A, B)$ can then simply be 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 (when dissimilarity is assumed to be related to a distance of judgements projected to an n–dimensional space) are not necessarily always valid in the perception [7, 6]. 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 $\widetilde{d}(A, B)$ correlated with $d(A, B)$ as highly as possible.

In this first attempt, we have chosen pitch-synchronous analysis of compared patterns, with microsegment of speech signal two pitch-periods long. We can define measurable distance between microsegments 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 microsegments in patterns A and B. Of course, as there is requirement for $\widetilde{d}(A, A) = 0$, it must be true that $\widetilde{d}_{i,j}(A, A) = 0, \forall i = j$. In the present paper we have chosen the following methods for $\widetilde{d}_{ij}$ computing.

Waveform dissimilarity 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 microsegments, cross-correlation defined by Equation (2) was applied on i, j pairs of von Hann window-weighted microsegments

$$\widetilde{d}_{i,j}(A, B) = 1 - \max_k \left(\frac{\sum_t^T A_i(t) B_j(t-k)}{\sqrt{\sum_t^T A_i^2(t) \sum_t^T B_j^2(t+k)}} \right) \tag{2}$$

where $t = 1, 2, \ldots, T$ and $k = 1, 2, \ldots, T$ are the indexes of samples in microsegments (of the same length). The dissimilarity of unvoiced segments was estimated simply by the ratio of the zero-crossing values of microsegments + their ratio of RMS, which led to values also in $\langle 0, 1 \rangle$ interval

$$\widetilde{d}_{i,j}(A, B) = 1 - 0.5 \frac{\sum_t^{T-1} \mathrm{sgn}(A_i(t) A_i(t+1))}{\sum_k^{K-1} \mathrm{sgn}(B_j(k) B_j(k+1))} - \\ - 0.5 \frac{\sqrt{\frac{1}{T_i} \sum_t^T A_i(t)^2}}{\sqrt{\frac{1}{T_j} \sum_k^K B_j(k)^2}} \tag{3}$$

where $\mathrm{sgn}(a) = 1$ iff $a < 0$ and $\mathrm{sgn}(a) = 0$ iff $a >= 0$. In the cases when $\widetilde{d}_{i,j}(A, B) > 1$ in (3), the new value is inverted $\widetilde{d}_{i,j}(A, B) = 1/\widetilde{d}_{i,j}(A, B)$. In the cases when voiced and unvoiced segments would have to be compared, cost 1 was returned immediately.

Singular value decomposition is an alternative approach employed in [12] for the measure of join cost. The author showed that the singular value decomposition (SVD) 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 (using microsegments) but global in scope, as all microsegments are decomposed using the same transform kernel. The SVD of signal matrix W is thus given by

$$W = USV^T \tag{4}$$

where W was created so that I microsegments of prosodic pattern A occupied rows $1, \ldots, I$, J microsegments of prosodic pattern B occupied rows $I+1, \ldots, I+J$ and so on, until all patterns of one prosodic word were added – let us, for this experiment, define function $\mathcal{F}(B, j)$ returning the index of row in W containing j^{th} microsegment of pattern B. The microsegments were von Hann's windowed signals, centred to row and surrounded by 0 if shorter than the number of columns in W (given by the longest microsegment). The dimension of SVD was here set to 10, as it was in [12]. The measurable distance was then computed as

$$\widetilde{d}_{i,j}(A, B) = 1 - \cos(u_k S, u_l S) = \\ = 1 - \frac{u_k S^2 u_l^T}{\| u_k S \| \, \| u_l S \|} \tag{5}$$

where $k = \mathcal{F}(A, i)$ and $l = \mathcal{F}(B, j)$.

To evaluate the measurable dissimilarity of the whole patterns, we have chosen symmetric DTW algorithm with slight modification ensuring that only microsegments from the same phones, with overlap to $1/3$

of the preceding and following phone allowed, are compared together (both compared patterns are always different realisations of one prosodic word).

For the given microsegment distance measure, the measurable dissimilarity is defined as

$$\widetilde{d}(A,B) = \min_{\{\mathcal{I}(k),\mathcal{J}(k),K\}} \left(\sum_{k}^{K} (\widetilde{d}_{\mathcal{I}(k),\mathcal{J}(k)}(A,B) * w_k) \right) \quad (6)$$

where $\mathcal{I}(k)$ and $\mathcal{J}(k)$ are functions warping k^{th} step in DTW into coordinates of compared microsegments 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 $w_k = 2$, $\mathcal{I}(k) = \mathcal{I}(k-1) \vee \mathcal{J}(k) = \mathcal{J}(k-1)$.

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 word. Measurable distance $\widetilde{d}(A,B)$ was also computed using the chosen measures between each of the same pairs. In Table 4 correlation of those two dissimilarities are summarised, together with the number of variants (patterns) being combined (for n variants, the correlation coefficient was computed from $n(n-1)/2$ pairs of $d(A,B)$ and $\widetilde{d}(A,B)$ values).

It can be seen that in this very first experiment, the correlations obtained are ranging from highly correlated to virtually uncorrelated (with even negative dependency in one case). 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. Instead, we assume that there are unsolved issues mainly in the perceived dissimilarity estimation, as summarised in Section 5.

Despite the variances in correlation, and the fact that 0.5 in average is not much, the most important finding is the validation that the whole idea, of which the (dis)similarity measure is a critical part, may be feasible.

Patterns no. + word		**Waveform**	**SVD**
6	potravin	0.653	0.539
6	pru:mislovi:x	0.936	0.879
6	spolupra:t_se	-0.416	-0.130
7	vminulosci	0.607	0.541
7	za:kazJi:ku:	0.265	0.422
7	nasvjece	0.834	0.576
8	nemu:Zeme	0.460	0.374
8	novina:P\u:m	0.718	0.727
9	pozornost	0.814	0.822
11	konkurent_se	0.577	0.277
11	t_Slovjeka	0.510	0.548
13	republit_se	0.242	0.291
14	proble:mu:	0.510	0.509
10	informat_si:	0.862	0.884
14	zdu:razJil	0.138	0.330
12	hospoda:P\stvi:	0.564	0.291
13	rospot_Stu	0.336	0.116
	Average:	0.507	0.470

Table 1: The correlation of perceived and measurable dissimilarities. The *words* are printed in SAMPA alphabet, each having *patterns no.* variants being combined together.

5. Conclusion

By no means did we aim to present a measure instantly applicable for perceived (dis)similarity estimation; we only intended to show the weak points of the current unit selection approach together with the proposal of possible solutions. We also attempted to show that the proposed approach is feasible, although there is still a large amount of work to do – yet we wish to avoid concluding that waveform dissimilarity is a better estimator than SVD.

First of all, we must pay extra attention to the results of listening tests used for the estimation of $d(A,B)$ – careful analysis and verification of listener responses aiming to determine unaccountable answers (the principle of listening tests does not consider any answer as "bad") is crucial, as relying on incorrect or otherwise distorted user responses must lead to biased conclusions in the evaluation of measurable dissimilarity metrics. Let us note, that the reliability of cross-participant agreement computed by means of Fleiss' kappa [13] is only 0.21, which does reject the null hypothesis that observed agreement is accidental on significance level 0.05, but does not make the agreement strong enough to make definite conclusions about how to measure perceptual similarity by acoustic signal. Currently, the answers in the listening test are reviewed, so it will be interesting to examine how the review shifts the kappa and the correlation computed in Table 4. Moreover, during building MDS dissimilarity matrix, there is also the possibility to employ the information from confusion matrixes [9] which contain likelihoods of evaluation confidence for individual listening test participants. The choice of the optimal dimension of MDS data representation must be taken into consideration as well.

We also plan to re–examine the proposed microsegment (dis)similarity measures as well as the use of DTW algorithm for the search of the best match. In addition, other ways of $\widetilde{d}(A, B)$ measure must 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 [6] or fuzzy feature contrast model [8] (it will, however, require the definition of predicates). After that, the correct behaviour of the best dissimilarity measure found should again be verified by means of listening tests, but for other voice(s).

6. Acknowledgement

This research is supported by the Grant Agency of the Czech Republic, project no. GACR 102/06/P205. Special thanks must also go to Mr. Greg Ashby (University of California, Santa Barbara) and Mr. Simone Santini (Universidad Autonoma de Madrid) for their valuable help with the explanation of ambiguities related to perceived similarity.

7. References

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

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

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

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

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

[7] 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.

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

[9] D. Tihelka, J. Romportl, "Statistical Evaluation of Reliability of Large Scale Listening Tests", submitted to ICSP 2008. Beijing, China, 2008.

[10] 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. Jeju, Korea, 2004.

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

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

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