Generalized N-gram measures for melodic similarity

Klaus Frieler

Institute for Systematic Musicology, University of Hamburg,
Neue Rabenstr. 13, D-20354 Hamburg, Germany

Abstract. In this paper we propose three generalizations of well-known N-gram approaches for measuring similarity of single-line melodies. In a former paper we compared around 50 similarity measures for melodies with empirical data from music psychological experiments. Similarity measures based on edit distances and N-grams always showed the best results for different contexts. This paper aims at a generalization of N-gram measures that can combine N-gram and other similarity measures in a highly general way.

1 Introduction

For similarity comparisons melodies are often viewed as sequences (strings) of pitch symbols. This is a quite natural approach in the light of common practice music notation and indeed proves to be quite sufficient and adequate for many applications. However, some important aspects of melodies such as the order of tones regarding to pitch height or the dimension of rhythm are often left out. In our former work (Müllensiefen & Frieler (2004)) we acheived optimized similarity measures as linear combinations of similarity measures coming from different musical dimensions such as pitch, contour, rhythm, and implicit tonality. We now aim at more complex combinations with a generalization of the N-gram approach. To do this we first to set out some basic concepts, viewing melodies as (time) series of arbitrary length in an arbitray event space. We go on with the definition of melodic similarity measures and present three common N-gram-based approaches for text similarity and their application to melodies. We will then develop a generalization of these N-gram measures using the concept of similarity on a lower level whereby we come close to some concepts of fuzzy logic.

2 Abstract melodies

The fundamental elements of symbolic music processing are discrete finite sequences. We will state some basic definitions at the beginning.

2.1 Discrete sequences

Definition 1 (sequences). Let $\mathbb{E}$ be a set and $N \leq M$ two integer numbers. A **sequence** over $\mathbb{E}$ is a discrete map

$$\begin{aligned}\phi : [N : M] &\to \mathbb{E}\\ k &\mapsto \phi(k)\end{aligned}$$

We write $|\phi| = M - N + 1$ for the length of a sequence. A sequence is said to have **normal form**, if $N = 0$. For sequences of length N we write $F_N(\mathbb{E})$. The space of all finite sequences over $\mathbb{E}$ is written as

$$F(\mathbb{E}) = \bigcup_{N \in \mathbb{N}_0} F_N(\mathbb{E}).$$

The set $F_0(\mathbb{E}) = \emptyset$ is the **empty sequence**.

Some technical remarks should to be made. First of all, the notation $[\cdot : \cdot]$ denotes a consecutive set of integer numbers, and we choose 0-base indexing for the normal form as it is common in the programming world (because as all this aims finally at some computer implementation). As it is clear the indexing of a sequence is not relevant, the index-shift operator establishes a equivalence relation on the space of sequences, and, unless otherwise stated, we will assume normal form for sequences throughout the following.

The N-gram approaches for similarity are based on the concept of subsequences length N (Downie (1999), Uitdenbogerd (2002)).

Definition 2 (N-gram). Let $s \in F_N(\mathbb{E})$ and $0 \leq i \leq j < N$. Then

$$\begin{aligned}\phi^i_j : [i : j] &\to \mathbb{E}\\ k &\mapsto \phi(k)\end{aligned}$$

is called a **N-gram** or **subsequence** of ϕ of length $n = j - i + 1$. The set of all N-grams of ϕ of length n is notated with $\mathcal{S}_n(\phi)$, the set of all subsequences with $\mathcal{S}(\phi)$.

Music can be abstracted and symbolically represented in many ways. Two main classes can be differentiated: audio (or signal) oriented and notation (or symbolical) oriented representations. For our purposes of measuring melodic similarity we use solely symbolically coded music on the abstraction level of common practice (western) music notation. Most music theorist and music psychologists agree that onset and pitch are mostly sufficient to capture the 'essence' of a melody. We share this viewpoint and will give now the definition of an abstract melody, considered as a finite, discrete sequence in some event space.

Definition 3 (Melody). Let $\mathbb{E}$ be an event space. A finite, discrete map

$$\begin{aligned}\mu : [0 : N-1] &\rightarrow \mathbb{R} \times \mathbb{E} \\ n &\mapsto (t_n, p_n),\end{aligned}$$

is called **melody** if

$$t_n < t_m \Leftrightarrow n < m.$$

The vaues p_n are called **generalized pitch**. The space of melodies of length N will be notated with $\mathcal{M}_N^{\mathbb{E}}$, the space of melodies of arbitrary (but finite length) with $\mathcal{M}^{\mathbb{E}}$.

3 Similarity measures

After this introduction of the basic notions, we will now discuss similarity measures for melodies. At the beginning we had to make the basic choice between the complementary concepts of similarity and dissimilarity (i.e. distance) measures. In the statistical practice the use of distance measures is far more common, e.g. for MDS or cluster analysis. However, our aim was to compare many different (dis-)similarity measures on a common ground and moreover with experimental data. Because the task to judge similarity is much easier and more familiar for musical experts than to judge dissimilarity, we choose similarity as our basic concept. Because of the complementary nature of the two approaches every similarity measure can be formulated as dissimilarity measure and vice versa, but unfortunately not uniquely, i.e. with some degree of freedom particularly due to choice of normalization and transformation function. But we will not delve into this topic too far here due to space limitations. Instead we state our definition of general similarity maps and of similarity measures for melodies.

Definition 4 (Similarity map). Let $\mathcal{M}$ be an arbitray set. A **similarity map** is a map

$$\begin{aligned}\sigma : \mathcal{M} \times \mathcal{M} &\rightarrow [0, 1] \\ (\mu, \mu') &\mapsto \sigma(\mu, \mu')\end{aligned}$$

with the following properties

1. Symmetry: $\sigma(\mu, \mu') = \sigma(\mu, \mu')$
2. Self-identity: $\sigma(\mu, \mu) = 1$ and $\sigma(\emptyset, \mu) = 0 \; \forall \mu \neq \emptyset$

The similarity map $\sigma(\mu, \mu') = 1$ is called the **trivial** similarity map. A similarity map with

$$\sigma(\mu, \mu') = 1 \Leftrightarrow \mu = \mu'$$

is called **definite**. The value of $\sigma(\mu, \mu')$ is the **degree of similarity** of μ and μ'.

A similarity map can be viewed as a generalization of Kroneckers δ-operator (which is of course a similarity map). Between the two distinct cases *identic* and *non-identic* a non-trivial similarity map provides a whole continuum of partial identity, i.e. similarity. This is related to concepts of fuzzy logic, where one has degrees of *belongingness* of elements to sets and degrees of *truth* for logical statements.

For similarity measures for melodies we demand some additional properties besides of being a similarity map:

Definition 5 (Melodic similarity measure). A **melodic similarity measure** is similarity map

$$\sigma : \mathcal{M}^{\mathbb{E}} \times \mathcal{M}^{\mathbb{E}} \to [0, 1]$$

which is invariant under pitch transposition, time shift and tempo change.

We will not go into the detail of the meaning of this additional properties, which are rooted in musicological findings, and refer the interested reader to Müllensiefen & Frieler (2004).

3.1 N-gram measures

N-gram based approaches form a set of standard techniques for measuring similarity of strings and texts and for abstract melodies as well (Downie (1999), Uitdenbogerd (2002)). All of them are more or less based on counting common N-grams in two strings. We will discuss here three basic forms, the Count-Distinct, the Sum-Common and the Ukkonen measure. For this and for later purposes we first define the frequency $f_s(r)$ of a N-gram r with respect to a sequence s:

$$f_s(r) = \sum_{u \in \mathcal{S}_n(s)} \delta(u, r)$$

(The δ-operator for sequences is obviously given through the identity of corresponding sequence elements.) The set of all distinct N-grams of a sequence s will be written as $\mathbf{n}(s)$

Definition 6 (Count-Distinct, Sum-Common and Ukkonen measure). Let s and t be two sequences over a event space $\mathbb{E}$, and let $0 < n \leq \min(|s|, |t|)$ be a positive integer.

1. The **Count-Distinct measure (CDM)** is the count of N-grams common to both sequences:

$$S_d(s, t) = \sum_{r \in \mathbf{n}(s) \cap \mathbf{n}(t)} 1 = |\mathbf{n}(s) \cap \mathbf{n}(t)|$$

2. The **Sum-Common measure (SCM)** is the sum of frequencies of N-grams common to both sequences

$$S_c(s,t) = \sum_{r \in \mathbf{n}(s) \cap \mathbf{n}(t)} f_s(r) + f_t(r),$$

3. The **Ukkonen measure (UM)** counts the absolute differences of frequencies of all distinct N-grams of both sequences:

$$S_u(s,t) = \sum_{r \in \mathbf{n}(s) \cup \mathbf{n}(t)} |f_s(r) - f_t(r)|$$

The Count-Distinct and the Sum-Common measures are (unnormalized) similarity measures, the Ukkonnen measures is a distance measure. To fulfill our definition of a similarity map we need to apply a normalization to the first and second, and for the UM we additionally need a suitable transformation to a similarity measure.

$$\sigma_d(s,t) = \frac{S_d(s,t)}{f(|\mathbf{n}(s)|, |\mathbf{n}(t)|)}$$
$$\sigma_c(s,t) = \frac{S_c(s,t)}{|s| + |t| - 2n + 2}$$
$$\sigma_u(s,t) = 1 - \frac{S_u(s,t)}{|s| + |t| - 2n + 2}$$

The choice of this normalizations come mainly from the self-identity property and from the desirable property of being definite. Natural choices for $f(x,y)$ are $\max(x,y)$ and $\frac{1}{2}(x+y)$. The normalization for the SCM comes from the cases of two identic constant sequences, and for the UM from the case of two constant sequences with no common N-grams. The choice of the functional form $1-x$ to transform the UM into a similarity map after normalization is arbitrary but quite natural and simple. Of course any monotonic decreasing function with $f(0) = 1$ and $\lim_{x\to\infty} f(x) = 0$ like e^{-x} could have been used as well.

Example 1. We will give an short example with 4-grams. We consider the beginning 6 notes of a major and a minor scale, e.g. in C we have $s = \{"C", "D", "E", "F", "G", "A"\}$ and $t = \{"C", "D", "Eb", "F", "G", "Ab"\}$. We neglect the rhythm component. We will transform this to a presentation using semitone intervals to acheive transposition invariance:

$$s = \{2,2,1,2,2\}, t = \{2,1,2,2,1\}$$

We have two 4-grams for each melody

$$s_1 = \{2,2,1,2\}, s_2 = \{2,1,2,2\}$$

$$t_1 = \{2, 1, 2, 2\}, s_2 = \{1, 2, 2, 1\}$$

with only one common 4-gram s_2 resp. t_1. Thus the normalized CD similarity is

$$\sigma_d = \frac{1}{2}$$

the normalized SCM is likewise

$$\sigma_c = \frac{1+1}{|s| + |t| - 2 \cdot 4 + 2} = \frac{1}{2}$$

and the normalized UM is

$$\sigma_u = 1 - \frac{1 + |1-1| + 1}{5 + 5 - 8 + 2} = \frac{1}{2}$$

Incidentally, all three measures give the same value for this example.

4 Generalized N-grams

4.1 Reformulation

We now come to the generalization procedure for the three N-gram measures. The basic idea is to generalize identity of N-grams to similarity. A common N-gram is a N-gram present in both sequences, or, stated in other words, the frequencies with respect to both sequences of a common N-gram is greater than zero. We will use this idea to restate the N-gram measures in a way more suitable for generalization.

For the following let be s and t be sequences over $\mathbb{E}$ of length $|s| = N$ and $|t| = M$, $0 < n \leq \min(N, M)$, and $0 \leq \epsilon < 1$ be an arbitrary real constant. Morever let σ be a similarity map for $F(\mathbb{E})$. A helpful function is the step function Θ which is defined as

$$\Theta(t) = \begin{cases} 1 & t > 0 \\ 0, & t \leq 0 \end{cases}$$

The **presence function** Θ_s of a N-gram r with respect to s can be defined as

$$\Theta_s(r) = \Theta(f_s(r) - \epsilon) = \begin{cases} 1 & r \in \mathcal{S}_n(s) \\ 0, & r \notin \mathcal{S}_n(s) \end{cases}$$

For compactness of presentation it is moreover useful to define the frequency ratio of a N-gram r with respect to s and t:

$$g_{st}(r) = \frac{f_s(r)}{f_t(r)}$$

We are now able to restate the basic definitions of the three N-gram measures.

Lemma 1. *The CDM, SCM and UM can be written as follows:*

1. *(CDM)*
$$S_d(s,t) = \sum_{r \in \mathcal{S}_n(s)} \frac{\Theta_t(r)}{f_s(r)}$$

2. *(SCM)*
$$S_c(s,t) = \sum_{r \in \mathcal{S}_n(s)} \Theta_t(r)(1 + g_{ts}(r))$$

3. *(UM)*
$$S_u(s,t) = \sum_{r \in \mathcal{S}_n(s)} 1 + \Theta_t(r)(\frac{1}{2}|1 - g_{ts}(r)| - 1) + (s \leftrightarrow t) \tag{1}$$

Proof. First, we note that a sum over distinct N-grams can be rewritten as sum over all N-grams:

$$\sum_{r \in \mathbf{n}(s)} F(r) = \sum_{r \in \mathcal{S}_n(s)} \frac{F(r)}{f_s(r)} \tag{2}$$

We will only proof the last formula for the Ukkonen measure, the other proofs follow similar lines of argumenation.

$$\begin{aligned}
\sum_{r \in \mathbf{n}(s) \cup \mathbf{n}(t)} |f_s(r) - f_t(r)| &= \sum_{r \in \mathcal{S}_n(s) \setminus \mathbf{n}(t)} 1 + \sum_{r \in \mathcal{S}_n(t) \setminus \mathbf{n}(s)} 1 \\
&\quad + \sum_{r \in \mathcal{S}_n(t) \cap \mathbf{n}(t)} \frac{|f_s(r) - f_t(r)|}{f_s(r)} \\
&= \sum_{r \in \mathcal{S}_n(s)} (1 - \Theta_t(r)) + \sum_{r \in \mathcal{S}_n(t)} (1 - \Theta_s(r)) \\
&\quad + \frac{1}{2} \left(\sum_{r \in \mathcal{S}_n(s)} |1 - g_{ts}(r)| \Theta_t(r) + (s \leftrightarrow t) \right) \\
&= \sum_{r \in \mathcal{S}_n(s)} 1 + \Theta_t(r)(\frac{1}{2}|1 - g_{ts}(r)| - 1) + (s \leftrightarrow t) \square
\end{aligned}$$

After this reformulation we again have to apply a normalization to fulfill our similarity map definition. We will state this without proof and define some auxiliary functions

$$\beta(s) = \sum_{r \in \mathcal{S}_n(s)} \frac{1}{f_s(r)},$$

$$\lambda(s,t) = \sum_{r \in \mathcal{S}_n(s)} 1 + g_{ts}(r) = \sum_{r \in \mathcal{S}_n(s)} \frac{f_s(r) + f_t(r)}{f_s(r)}$$

and

$$\eta(s,t) = \sum_{r \in \mathcal{S}_n(s)} 2 + |1 - g_{ts}(r)|$$

With this functions we can write down the normalized similarity measures in a compact way:

$$\sigma_d(s,t) = \frac{S_d(s,t)}{f(\beta(s), \beta(t))}$$
$$\sigma_c(s,t) = \frac{S_c(s,t)}{f(\lambda(s,t), \lambda(t,s))}$$
$$\sigma_u(s,t) = 1 - \frac{S_u(s,t)}{f(\eta(s,t), \eta(t,s))}$$

Again the function $f(x,y)$ can be $\max(x,y)$ or $\frac{1}{2}(x+y)$. For the following we fix f to the second form of the arithmetic mean.

4.2 Generalization

This new form of the N-gram measures can now be generalized. The first step is the introduction of generalized frequencies:

$$\nu_s(r) = \sum_{u \in \mathcal{S}_n(s)} \sigma(u,r) \geq f_s(r)$$

and generalized frequency ratios:

$$\omega_{st}(r) = \frac{\nu_s(r)}{\nu_t(r)}$$

By substituting frequencies with generalized frequencies the desired generalization is achieved. Now arbitrary similarity maps (and measures) can be used to judge the degree of similarity between N-grams. The old definition is contained as a special case of this generalized N-gram measure with Kronecker's δ as similarity map.

Example 2. We take the same two melodies as in example 1 with 4-grams and use a similarity measure based on the Levensthein-Distance $d(u,v)$ (see e.g. Müllensiefen & Frieler (2004) for a definition and discussion):

$$\sigma(u,v) = 1 - \frac{d(u,v)}{\max(|u|, |v|)}$$

First of all we need the similarities between all 4-grams:

$$\sigma(s_1, s_2) = \sigma(s_1, t_1) = \sigma(s_2, t_2) = \frac{1}{2}$$
$$\sigma(s_1, t_2) = 0$$

and calculate the generalized frequencies:

$$\nu_s(s_1) = \nu_s(s_2) = \nu_t(t_1) = \nu_t(t_2) = \nu_t(s_2) = \nu_s(t_1) = \frac{3}{2}$$
$$\nu_t(s_1) = \nu_s(t_2) = \frac{1}{2}$$

with generalized frequency ratios:

$$\omega_{st}(t_1) = \omega_{ts}(s_2) = 1$$
$$\omega_{st}(t_2) = \omega_{ts}(s_1) = \frac{1}{3}$$

Now we determine the presence functions of all 4-grams with $\epsilon = \frac{1}{2}$.

$$\Theta_t(s_1) = \Theta_s(t_2) = \Theta(\frac{1}{2} - \frac{1}{2}) = 0$$
$$\Theta_t(s_2) = \Theta_s(t_1) = \Theta(\frac{3}{2} - \frac{1}{2}) = 1$$

As a last step we determine values of the auxiliary functions:

$$\beta(s) = \beta(t) = \frac{4}{3}$$
$$\lambda(s,t) = \lambda(t,s) = \frac{10}{3}$$
$$\eta(s,t) = \eta(t,s) = \frac{14}{3}$$

With this preleminaries we can now calculate the three generalized N-grams measures. First the GCDM:

$$\begin{aligned}\sigma_d(s,t) &= \frac{2}{\beta(s)+\beta(t)} \sum_{r \in \mathcal{S}_n(s)} \frac{\Theta_t(r)}{\nu_s(r)} \\ &= \frac{2}{4/3+4/3} \frac{1}{\nu_s(s_2)} \\ &= \frac{1}{2}\end{aligned}$$

Then the GSCM:

$$\begin{aligned}\sigma_c(s,t) &= \frac{2}{\lambda(s,t)+\lambda(t,s)} \sum_{r \in \mathcal{S}_n(s)} \Theta_t(r)(1+\omega_{ts}(r)) \\ &= \frac{3}{10}(1+\omega_{ts}(s_2)) = \frac{3}{5},\end{aligned}$$

and at last the GUM:

$$\begin{aligned}\sigma_u(s,t) &= 1 - \frac{2}{\eta(s,t)+\eta(t,s)} \sum_{r\in\mathcal{S}_n(s)} 1 + \Theta_t(r)(\frac{1}{2}|1-\omega_{ts}(r)| - 1) + (s \leftrightarrow t)\\ &= 1 - \frac{2}{14/3+14/3}(1+\frac{1}{2}|1-\omega_{ts}(s_2)| + 1 + \frac{1}{2}|1-\omega_{st}(t_1)|)\\ &= 1 - \frac{3}{14}(1+1) = \frac{4}{7}\end{aligned}$$

We see from this example that the generalized measures usually raise the similarity values compared to the orginal version.

There is some more possibility for further generalization. For this purpose we define a ramp function:

$$\rho(t) = \begin{cases} 0, \, t \leq 0 \\ t, \; 0 \leq t < 1 \\ 1, \, t \geq 1 \end{cases}$$

and can now generalize the presence function (with some real constant $a > 0$):

$$\theta_s(r) = \rho(\frac{\nu_s(r)}{a})$$

Example 3. We consider again the above example and will calculate the GCDM with this new generalized presence function ($a = 1$).

$$\begin{aligned}\theta_t(s_1) &= \rho(\nu_t(s_1)) = \rho(\frac{1}{2}) = \frac{1}{2}\\ \theta_t(s_2) &= \rho(\nu_t(s_2)) = \rho(\frac{3}{2}) = 1\\ \theta_s(t_1) &= \rho(\nu_s(t_1)) = \rho(\frac{3}{2}) = 1\\ \theta_s(t_2) &= \rho(\nu_s(t_2)) = \rho(\frac{1}{2}) = \frac{1}{2}\end{aligned}$$

Applying this to the GCDM gives the following similarity value for our example.

$$\begin{aligned}\sigma_d(s,t) &= \frac{2}{\beta(s)+\beta(t)} \sum_{r\in\mathcal{S}_n(s)} \frac{\theta_t(r)}{\nu_s(r)}\\ &= \frac{2}{4/3+4/3}(\frac{\theta_t(s_1)}{\nu_s(s_1)} + \frac{\theta_t(s_2)}{\nu_s(s_2)})\\ &= \frac{3}{4}(\frac{1/2}{3/2} + \frac{1}{3/2}) = \frac{3}{4}\end{aligned}$$

5 Conclusion

We proposed a generalization of well-known similarity measures based on N-grams. The application of these techniques to melodies made a generalization desirable because of the cognitively multidimensional nature of melodies. They can be viewed to some extent as string of symbols but this already neglects such important dimension as rhythm and pitch order. Besides the possibility of using linear combination of similarity measures that focus on different musical dimensions, it could be fruitful to combine this measure in a more compact way. This, however, waits for further research, particularly an implementation of the generalized N-gram measures proposed here (which is currently under development), and a comparision with existing empirical data and other similarity measures. Thus, this paper should be viewed as a first sketch of ideas in this direction, merely a proof of concept.

References

DOWNIE, J. S. (1999): *Evaluating a Simple Approach to Musical Information retrieval: Conceiving Melodic N-grams as Text.* PhD thesis, University of Western Ontario .

MÜLLENSIEFEN, D. & FRIELER, K.(2004): Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments. *Computing in Musicology*, Vol. 13.

UITDENBOGERD, A. L. (2002): *Music Information Retrieval Technology.* PhD thesis, RMIT University Melbourne Victoria, Australia.