# A Full-body Gesture Recognition System and its Integration in the Composition "3rd. Pole"

*Miha Ciglar*
University of Music and
Dramatic Arts Graz, Austria
++386 40 512 603
miha.ciglar1@guest.arnes.si

**ABSTRACT**

*"3rd. Pole"* is the working title for a composition that is to be performed by a dancer on a specially designed gestural interface, based on optical motion tracking technology. This paper introduces the technical and artistic concepts developed during the last one year of research. Though some of the artistic ideas were already introduced in my earlier works [7], they are tightly interwoven with the technical concepts of pattern recognition techniques, which are clearly dominating my current research and the content of this paper. It is important that the reader is aware of the fact that the motivation behind this project is inspired by a new approach to composition, with the goal of generating a musical result. However, the concepts of full-body gesture recognition that are to be described in the first chapters may be repurposed or integrated in any other practical or artistic context. The composition "3rd. Pole" is still in development and has not publicly been performed yet, hence there is always the possibility of conceptual deviations when all the elements described in this paper will eventually come together in the final realization.

## 1. TECHNICAL DESCRIPTION

### 1.1 Infrastructure

It is convenient to define the whole system as an instrument. As common to all instruments, also this one has an input and an output section. The input section is a *"Vicon 8"* motion capture system [13] monitoring the dancer's actions. A brief description of the system and some common data mapping strategies for musical applications can be found in [2]. The *"Vicon 8"* system consists of 12 infra red cameras / sensors, placed around the dancer (the performance area). It is able to track and extract the Cartesian x/y/z coordinates of light-reflecting markers on his body in 3 dimensional space, at a sampling-rate of up to 120 frames per second. In order to start as simple as possible, only 4 points of the human body were tracked. The markers were arranged in groups, so that a characteristic constellation of 4 to 5 markers attached to the end of each limb (fig.1) would represent one central point from which we received our spatial coordinates. The trajectories of those coordinates were then used as an input for a gesture recognition algorithm, implemented in the real-time programming environment: PD (Pure data) [11], which is receiving the location data from the Vicon server through the OSC communication protocol [14]. There are two output sections on this instrument. As usual, there is an output in form of sound, and what is more, a second output in form of electricity, applied directly to the dancer's body, through a cable he is holding in his mouth. Ideally the dancer should be equipped with a wireless system, in order to avoid getting tangled in the cable. The electrical signal is another instance of the audio output itself, and is causing a waveform (amplitude) dependant sensation of pain. The purpose of this setup will be explained in detail later in the text.

**Fig.1:** markers attached to the dancer's limbs

### 1.2 Gesture Recognition

#### *1.2.1 Definition of a gesture*

When browsing through literature one finds several different definitions and notions of a gesture, especially in the work about gesture controlled instruments or gestures in performing arts. A comprehensive analysis of the terminology, relevant to the research field of gestures in musical contexts can be found in [5]. The notion of a full-body gesture referred to in this paper, is a short choreographical sequence with a clearly defined beginning and an end. Unlike other concepts that define gesture rather as a *style* of executing arbitrary *content* –

like in [6], where different emotional expressions like anger, joy, etc. are being tried to classify on the basis of features like "fluency" or "impulsiveness" which are interpreted from raw sensor data – recognizing specific *content*, regardless of the *style* of its execution is the focus of my work. A predefined (recorded) sequence of temporal relations among selected bodily features (derived from sensor values) should characterize the *content* of a gesture. It should further be possible to identify the same relational progression in the recognition process, whilst allowing a certain degree of deviation from the recorded example. Perhaps it would make sense to distinguish between *style* (a manner of performing an action) and *content* (the action itself) of a gesture. Both components (communication channels) are of equal importance and both have a potential capacity for carrying an equivalent amount of information. With respect to this subdivision, only the *content* component is addressed in this work.

### *1.2.2 Related work*

There are freely available tools for convenient multidimensional signal processing, like the MNM and FTM libraries [4] developed at IRCAM, and can be used within the MAX/MSP programming environment [10]. The gesture follower algorithm presented in [3] is implemented with these tools. Its code is open – (within the MAX/MSP platform) and is included as an example in the FTM library. It is based on the concept of left-to-right Hidden Markov Models [12] and is able to follow the progress of a predefined gesture in real-time. The algorithm operates with velocity values (first derivatives of the input data), (re-) sampled at regular time-intervals. I did not get to test it with multidimensional data, but the results achieved with the two-dimensional drawing interface in the example were promising. The problem though in using unprocessed velocity values, is that it is not possible to identify a gesture if its velocity or the duration of its execution differs from the recorded example.

### *1.2.3 Previous work – (pre-processing)*

The algorithm proposed in this paper is based on the results and observations of previous work. The ambition presented in [8] was to make the system unsusceptible to variations of *intensity* and *temporal evolution* of a gesture. *Intensity* would refer to the spatial extent in the execution of a predefined gesture, whereas an unrestricted *temporal evolution* enables a free, interpretation of gesture progression in time. Even a variation in the proportions of individual gesture segments should be possible. Further, the recognition system would not depend on the orientation of the dancer in space, allowing a correct recognition also if a gesture is performed lying on the floor and compared against its version in the upright position. The orientation independency is achieved, by taking the inter-point dynamics as the input parameters for the recognition algorithm. The Euclidian distances between each pair the four body points are constantly being measured and observed. With four points we get six distances (variables) to operate with, whereby we also achieve a dimension reduction of the original 12 dimensional data (4 points * 3 coordinates). Next, the first derivative (velocity) of the distance variations is calculated and quantized to 3 possible states: "1" for increasing, "-1" for decreasing and "0" for a constant distance. This quantization should allow for variations in the velocity (speed) of execution. So, our pre-processed input parameter is a 6 dimensional vector with its individual dimensions confined to 3 discrete states, which amounts to a total of $3^6 = 729$ possible states. This state vector is continuously being monitored, but not sampled at regular intervals; instead, it is only recorded each time one of its dimensions would change its value, thus discarding all information about absolute timing. Having no information about the exact sequence of velocity values but merely the information on the sequence of significant changes – which is also a justified interpretation when considering the human perception of a gesture and the abstraction of particular realizations to a common representation – the dancer is free to conduct a gesture with an arbitrary spatial extent and is also not bound to follow the exact temporal proportions of different gesture segments. Fig.2 shows a three dimensional example of how the state vectors are being generated. The continuous sequence of velocity values can be represented (generalized) by 9 state vectors, starting with: [0;0;0], [-1;1;0], [-1;1;1] and so on.

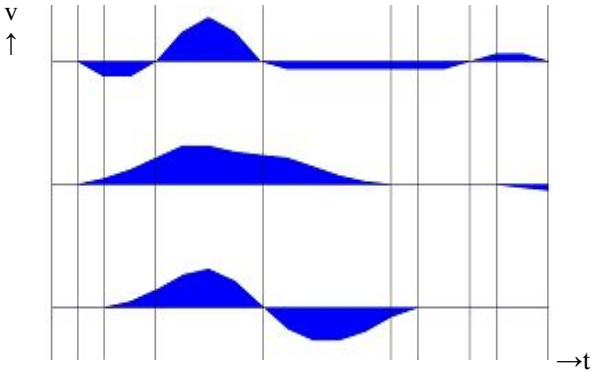


**Fig.2:** velocity graphs and discrete state change locations

### *1.2.4 Problems with recognition*

In the recognition phase, the incoming signal of the performed gesture was compared against a set of recorded gestures, starting with the first state of each recorded example, continuing with the next once a match would occur and returning to comparing the first state, once the states mismatched. There was also a time warping function that allowed jumping back and forth in a certain neighborhood of the currently compared state, before braking up the comparison due to a

mismatch. The problem with this approach however, was that minor and trivial movements could generate a large amount of data in short time, which significantly increased the chances of a false recognition. Moreover, correct recognition was often obstructed due to a slightly altered succession of states following each other in very short time. It was observed that the state vectors were generally generated burst-wise. For example: if only one limb is active (moving), a change of its course would most likely lead to a simultaneous change of the values in three dimensions (the distances to the other 3 static limbs) – fig. 3. Perceptually, the change happens simultaneously, but, at a sampling rate of 120 Hz, the measurement would rarely return more than one value at a time, but what is more, the specific succession of states is subject to variation in different realizations. Therefore, it is not possible to foretell the exact sequence of local changes, and a mechanism needed to be installed to thwart this phenomenon.

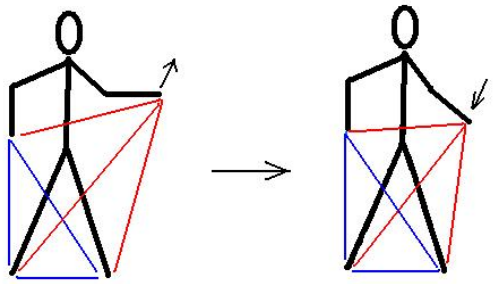

**Fig.3:** one active limb and the consequential change of three distances

### *1.2.5 Gesture segmentation*

The subdivision of a gesture into smaller segments, containing several state vectors and a group analysis seemed to be a good solution. In [1], the authors introduce a segmentation method together with the term *atomic gestures,* referring to "those that cannot be further decomposed, and which can be combined to create larger composite gestures". Those atomic segments are defined by the number of peaks in the second derivatives (acceleration) of a one dimensional data stream. Due to the multidimensionality of data in my approach the segmentation cues were not derived from specific patterns of acceleration values, although we can detect some interesting correlation between segment boundaries and the valleys in the acceleration traces shown in fig. 4. Since it was observed that the state vectors are generated burst-wise and, like in the above mentioned example, that those temporal clusters of state vectors actually mark the transition points (boundaries) of gesture segments, we are not comparing actual segments but the sequence of segment transition points, which contain most of the relevant information. Although these clusters are generated in a very short time, they include all the directional information of the preceding as well as the following segment. The clusters are defined (separated) by monitoring the time intervals between successive state vectors – fig. 5. A time limit is set, beyond which a new state vector is assigned to the new (next) cluster. Now, instead of analyzing the exact progression of state vectors, we analyze the sequence of state clusters, disregarding the order of individual state vectors inside a cluster. With this strategy, the intra-gesture tolerance, as well as the inter-gesture discrimination was improved. The recognition algorithm got less susceptible to error due to slight (local) state sequence variations in different realizations, with very little drawback in terms of fluentness of following a gesture (compared to the original approach). Also, a misinterpretation of a gesture got les likely, since now there is always a cluster of states needed to be compared in order to certify a match.

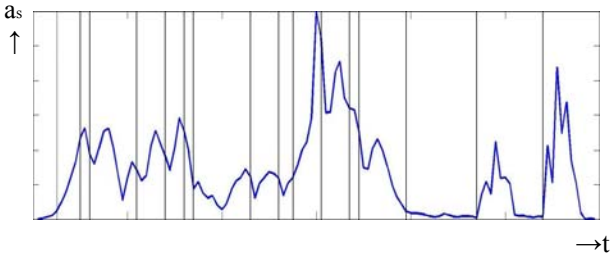


**Fig.4:** cluster locations compared to the absolute sum of all (n=6) acceleration values ( $a_s(t) = \sum_n |\partial^2 x_n / \partial t^2|$ )

### *1.2.6 Realization*

Handling multidimensional data inside PD is not very convenient, and the PD community is all looking forward to working with a tool like FTM [4]. According to [15] it might actually be available inside PD in the near future, but for now, every dimension needs to be taken care of "manually", by assigning special buffers to each individual dimension as well as complex buffer structures to each prerecorded gesture that is to be recognized. Consequentially, each gesture is defined within an autonomous module. We simply create any number of identicall modules – according to the amount of gestures we want to recognize. All of them are monitoring the (live) input data stream, and following their contained gesture without any knowledge regarding the activity of other gesture modules (models). The recorded data is segmented into clusters, the amount of which defines the relative length of a gesture. The clusters can only be identified as a whole, so they represent the smallest unit of the recognized gesture. It would often occur that a state is found several times inside a single cluster; therefore the cluster buffer was extended with an extra dimension, containing the states frequencies of appearance (their priority) in the given cluster. The higher this frequency, the grater the importance of a state, which can further be defined as "characteristic" to a particular cluster. In the comparison stage, the vectors are weighted with respect to their frequency of appearance. Depending on the spatial resolution of our optical sensors, more or less states are being generated at a unit of time. In fig. 2, for example, the resolution is kept relatively low, in order to stabilize

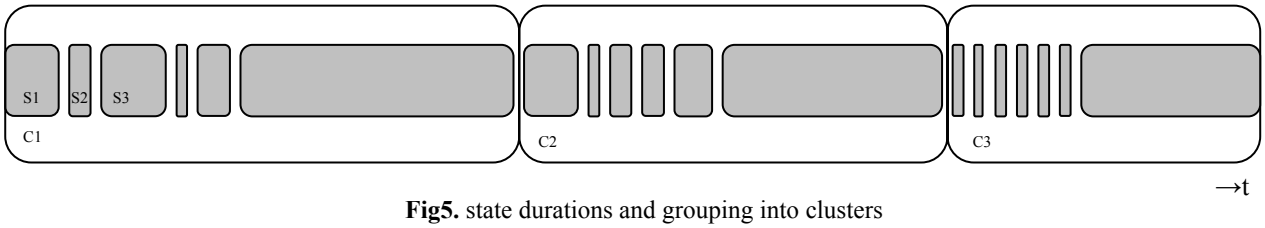


**Fig5.** state durations and grouping into clusters

the "zero state", and consequentially the states would not reoccur very often in a cluster. The recorded cluster is then compared against a cluster, generated in real time, and a percentage of correct state-vector matches – with respect to the largest common cluster size – is returned. The recognition step outputs a degree of completion of a particular gesture (in %) and returns to monitor the first cluster, after the full gesture is recognized or after a mismatch is detected.

*1.2.7 Results and observations*

Several different variations in the fine-tuning of the described concept were examined, and the recognition rate has noticeably been improved. The algorithm was not extensively tested yet and a wider range of different data would be needed to verify its functionality. I worked with a set of nine different and multiply recorded full-body gestures, which were taken from the classical ballet repertoire. The false recognition rate is still a bit problematic, since a false gesture interpretation may reach up to 30 % before a mismatch would appear and brake up the comparison process. Perhaps, those mistakes are unavoidable, considering the low dimensionality of our data. The trajectories of four body points can not accurately model complex full body gestures, and an improvement could definitely be achieved by increasing the number of tracked points. The main problem with the temporal clustering approach however, is, that a variation in the execution time of a gesture would not return the expected results (a correct recognition). This problem might be overcome by not selecting an absolute threshold for the clustering condition, but making it adapt in real-time to the activity (dynamics) of the input data. Perhaps the incorporation of a similar concept like the "activity detection function" presented in [1] might lead to the desired results.

### 1.3 Feedback

By moving through space, the dancer conducts actions in three spatial dimensions plus one temporal dimension. A fundamental part of the musical composition is the function that translates those actions to a two dimensional space (a time varying amplitude (the audio signal)), and will undergo a detailed discussion later in the text. The dimension of amplitude refers to the (fast changing) electronic signal waveform, corresponding to the sound being generated and projected. In addition to the sonification of the electronic waveform, which produces an auditory feedback, the dancer is also exposed to an alternative instance of the same signal. This instance is the (amplified) signal itself, in its primary (the electronic) domain. The connection with the dancer is established by a cable, which he is holding in his mouth. This concept of direct electronic signal-feedback was already applied and discussed in my earlier compositions and interface designs [9]. It enables the dancer / performer to experience an alternative impression of the induced sound. Since it is electricity we are dealing with here, the dancer would feel a pain with waveform (sound amplitude) dependant intensity. Therefore, we need to be very careful with the amplification of the signal in order not to seriously harm the dancer.

## 2. ARTISTIC CONCEPTION

There are a few parallel conceptions interweaved to form the framework of this composition and need to be discussed separately. One of the main intentions of this project was to blur the causal relationship of movement and sound, in order not to expose the spectacular potential of the employed technology in the first place. However, the approach of generating musical parameters via gestural cues should not restrain the control data to discrete values emerging at the end of a successful completion of a predefined gesture. The continuous progression of the gesture should be mirrored in a continuous change of sonic parameters, while responding merely to a specific selection of choreographic content.

### 2.1 Instrument Design

Like the majority of my recent work [7], *"3rd. Pole"* was developed as an experimental realization, accompanying the design of a new interface for musical expression. The basic idea was to have a musical instrument, which would generate sonic output according to the way it is being played. Musical patterns should be analyzed and reduced to their structural characteristics, with the element of "variation" defined as a primal feature. Patterns with different content but same characteristics of variation – for example: a trill played on an arbitrary note (pitch) – would be recognized and defined as identical. However, this primary instrumental output used for analysis should not be audible but should only be used as a control parameter for generation of the actual (desired) audio output in electronic form. It would be interesting to

combine the (primary) output of a traditional musical instrument with this concept and it will certainly be a challenging objective for some future work, but at this point, the focus was on finding an "instrument" without a primary output, to start with, and the idea to employ a dancer seemed quite intriguing. In this sense, the dancers function is re-contextualized, since we are only exploiting his ability of virtuosic control over his body, with the initial intent of generating musically interesting and diverse material. The dancer should therefore be viewed as a musician in the first place, and it is important that the audience is made aware of the fact that the piece is not necessarily a dance performance, like it appears to be at first glance. We can actually find interesting correlations between a dancer and a musician, since both of them are able to control and change the shape of their body with a high degree of virtuosity and accuracy. A musician would use a mechanical instrument to transform his bodily actions to the sonic domain, whereas a dancer does not have this interface and conducts his actions merely in the visual domain. With a motion tracking system, we can define a 3D space as an invisible instrument, allowing the musician (dancer) to perform the "trills", or any other gestural patterns anywhere in space.

### 2.2 Algorithmic Composition

The notion of algorithmic composition has many different facets, among which, the aspect of composing aurally unthinkable music is most alluring to me. In this case, the process of composition is confined to creating an alternative (a *secondary*) domain – that might be defined even without the dimension of time (which on the other hand is constitutive for the sonic domain) – and the arrangement of idiosyncratic elements of this particular parameter-space (domain) to a structure, defining the formal fundament of the final composition. Furthermore we need to "compose" a transfer function, in order to transform the composed events from the *secondary* to our primary (the sonic) domain, in which our desired (the audible) results would finally come to the fore. The earlier mentioned concept of electronic-signal feedback (sound in electric form applied directly to the body of the performer) is an instantaneous forward-backward implementation of an algorithmic composition. Here, the *secondary* domain is defined by the architecture of the instrument, the algorithms for sound generation and especially the feedback concept. The "forward" transfer function (Hf) is a subjective gesture to sound mapping strategy whereas the "backward" transfer function (Hb) is simply just the audio signal in its original – the electronic – form and can actually be interpreted as a variable inside Hf. The definition of compositional events is thereby possible only in real-time, since the *secondary* domain of this particular composition would by its nature require an immediate response to an electronic impulse the musician (dancer) has just generated. A (temporal) change of body-posture (= a gesture in 3 dimensional space and time), would generate an audible and in parallel a haptic (electrical) signal as a consequence (1 dimension + time). These in turn would generate a further dislocation of the dancer's body-posture (3 dimensions + time), where the electrical signal in particular would stand as a primary reference and as a factor with highest priority (as opposed to the sonic signal that stands merely as an aesthetic argument) for the dancer's decision on the content of the bodily actions he is going to undertake in future, for those would often be induced by an instantaneous physical reaction to the electric shock the dancer is exposed to.

### 2.3 Music & Choreography

A traditional method of how to synchronize dance and music in a non improvisatory sense would be: a dancer following the prerecorded musical score. In this project we wanted to have a complex musical composition, with several temporal irregularities, that would make it difficult for the dancer to follow its progression. On the other hand we also wanted to place an equal amount of importance on choreographic as well as musical elements, but what is more, to put the highest priority on the relative temporal stability of the "polyphonic" enmeshment of movement and sound. The individual gestures of the choreography should always be supported by an identical musical structure, so the dancer would not need to worry about the temporal accuracy of his choreography. The selected sounds would follow his actions (gestural cues) and not vice versa, however, without a bluntly noticeable synchronization effect – resulting from direct mappings strategies of location data to musical parameters. All the technology needed to realize this idea is on the stage and is visible but its functionality should remain invisible to the spectator. The audience should just enjoy a dance performance, without being overwhelmed by fascination over the potentially spectacular expression capabilities of the motion tracking technology.

### 2.4 3rd. Pole

In a dance performance, there are usually 2 elements (visual and audible) that need to be arranged and put into a contrasting or harmonizing etc. context. The title *"3rd. Pole"* should indicate the inclusion of a third, a haptic component contributed by the electronic current running through the dancer's body. He is exposed to a situation where he is in absolute decision power and needs to consider and outbalance all three elements (poles). Like already mentioned, we have the induced sound respectively its electronic abstraction, which is in direct contact with the performer's body. This enables a different corporal perception and interpretation of the caused sound, since now the performer does not only have the audible but also a haptic reference - i.e. pain, caused by the electric current - for the choice of his following actions. Therefore, also the process of

composition or better to say, the final arrangement of pre-composed material is only possible in real time, since we are interested in an alternative arrangement of the choreographic and musical progression, which is inspired by all three "poles" together. A pre-composed form or sequence of events would not make any sense, apart from satisfying possible sadistic tendencies of the composer.

## 3. CONCLUSION AND FUTURE WORK

In this paper, a system for recognizing full body gestures - from the interpretation of relatively few parameters - was introduced. Further, a practical example (the composition *3rd.Pole*) deploying the proposed algorithms was discussed and analyzed. A few parallel artistic conceptions were introduced that perhaps might appear mutually exclusive, but at the same time, the richness of interpretations can provide a multitude of focus areas for future work. The algorithms for gesture recognition will need to be improved in order to achieve the desired goal of tolerating variations in duration and velocity of different gesture realizations. In the next phase, also the amount of tracked body points will be increased, which should further increase the inter-gesture discrimination ability.

## 4. ACKNOWLEDGMENTS

This project has been supported by STEIM – www.steim.org – in Amsterdam (NL), by offering me a research residency in 2007; further, the Institute for Electronic Music (IEM) – www.iem.at – in Graz (A), by providing the facilities and technical infrastructure. Special Thanks goes to Maja Arzenšek for dancing and choreographical suggestions, Dr. Gerhard Eckel and David Pirro for providing theoretical opinions, IOhannes Zmölnig for technical assistance.

## 5. REFERENCES

[1] Benbasat, A, Y. and Paradiso, J. A. An inertial measurement framework for gesture recognition and applications. In Proceedings of the International Gesture Workshop, p. 9 – 20, London, UK, April 2001

[2] Bevilacqua, F., Dobrian, C. "Gestural Control of Music Using the Vicon 8 Motion Capture System", Proc. of the International Conference on New Interfaces for Musical Expression (NIME 03), Montreal, Canada, 2003

[3] Bevilacqua, F., Fléty, E., Guédy, F., Leroy, N., Schnell, N. "Wireless sensor interface and gesture-follower for music pedagogy", Proc. of the International Conference on New Interfaces for Musical Expression (NIME 07), New York, NY, USA, 200

[4] Bevilacqua, F., Muller, R., Schnell, N. "MnM: a Max/MSP mapping toolbox ", Proc. of the International Conference on New Interfaces for Musical Expression (NIME 05), Vancouver, Canada, 2005.

[5] Cadoz, C. and Wanderley, M. M. Gesture-Music. In M. Wanderley and M. Battier (eds.) Trends in Gestural Control of Music". Paris, Fr: IRCAM - Centre Pompidou, 2000

[6] Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R., Volpe, G. "Multimodal analysis of expressive gesture in music and dance performances", in A. Camurri, G. Volpe (Eds.), Gesture-based Communication in Human-Computer Interaction LNAI 2915, Springer Verlag, 2004

[7] Ciglar, M. homepage: www.ciglar.mur.at

[8] Ciglar, M. "3rd. Pole – a composition performed via gestural cues", in Proc. of the International Conference on New Interfaces for Musical Expression (NIME 08), Genova, Italy

[9] Ciglar, M. "I.B.R. Variation III." Proceedings of the EMS – Electroacoustic music Studies Network Conference – Beijing, China - October 2006

[10] Max/MSP programming environment www.cycling74.com/products/maxmsp.html

[11] Puckette, M. "Pure Data" Proceedings of the ICMC, 1996

[12] Rabiner, L. R. and Juang, B. H., "An introduction to hidden Markov models," IEEE Acoust. Speech Sign. Process. Mag. 3 (1986) 4-16.

[13] Vicon 8 motion capture system: www.vicon.com/entertainment/technology/v8

[14] Wright, M. "Open Sound Control: an enabling technology for musical networking" Organised Sound, 2005/12/01, Volume 10, Issue 3, p.193-200, (2005)

[15] Zmölnig, M., IO., Musil, T., Ritsch, W., Schnell, N. Freer than Max – porting FTM to Pure Data, in Proceedings of the Linux Audio Conference, Köln, 2008.