FP7-ICT-2013-C FET-Future Emerging Technologies-618067

SkAT-VG: Sketching Audio Technologies using Vocalizations and Gestures

D3.3.2
Final comprehensive annotation of the database of imitations

First Author	Pétur Helgason
Responsible Partner	KTH
Status-Version:	Final-0.1
Date:	December 29, 2015
EC Distribution:	Consortium
Project Number:	618067
Project Title:	Sketching Audio Technologies using Vocalizations and Gestures

Title of Deliverable:	Final comprehensive annotation of the database of imitations
Date of delivery to the EC:	31/12/2015

Workpackage responsible for the Deliverable	WP3
Editor(s):	Davide A. Mauro
Contributor(s):	Pétur Helgason
Reviewer(s):	Davide A. Mauro, Davide Rocchesso
Approved by:	All Partners

Abstract	The current deliverable presents the results of tasks T3.1 and T3.2.
Keyword List:	annotation database

The SkAT-VG Consortium consists of the following entities:

#	Participant Name	Short-Name	Role	Country
1	Università Iuav di Venezia	IUAV	Co-ordinator	Italy
2	Institut de Recherche et de Coordination Acoustique/Musique	IRCAM	Contractor	France
3	Kungliga Tekniska Högskolan	KTH	Contractor	Sweden
4	Genesis SA	GENESIS	Contractor	France

Document Revision History

Version	Date	Description	Author
First draft	16/07/2015	Import from .tex template	DAM
First draft	09/10/2015	Main contributions	PeH
First draft	16/07/2015	Minor reviews	DAM
Draft	15/10/2015	For reviewers	DAM

Table of Contents

Index of Figures

List of Acronyms and Abbreviations

DoW Description of Work

EC European Commission

PM Person Months

WP Work Package

1 Overview

The SkAT-VG articulatory database in Deliverable D3.3.2 contains annotated imitation data for eight imitators. The data are divided into two subsets, four professional actors (two male and two female) recorded at KTH, as well as four laypersons (two male and two female) recorded at IRCAM. The KTH data comprise 200 imitations, 50 for each imitator, and the IRCAM data comprise 208 imitations, 52 for each imitator.

The recording procedures for the KTH dataset were described in Deliverable D2.2.2. For the IRCAM dataset, the recording procedures were described in sections 2 and 3 of Deliverable D4.4.1.

Both datasets have been labelled using a set of 8 articulatory parameters. A detailed description of the annotation system was given in Deliverable D3.3.1.

2 ELAN - The database tool

For annotation and data extraction of articulatory data, SkAT-VG uses ELAN (EUDICO Linguistic Annotator), a freeware database tool produced by The Language Archive project at the Max Planck Institute for Psycholinguistics, Nijmegen, NL. ELAN is an annotation tool for creating, editing and extracting data from multi-media recordings. It is specifically designed for the analysis of language and gesture, but can be utilized for the annotation, analysis and documentation of any video and/or audio data.

3 Source files

3.1 The KTH data set

The KTH dataset was recorded with a focus on articulatory analysis. The media sources comprise 2 video recordings, an audio recording and an electroglottographic (EGG) recording. In order to preserve the context in which imitations were performed (preceding and following trials), entire sessions were recorded (rather than individual imitations). There are normally 5 sessions per imitator, comprising 10 imitations each. Sessions vary in duration, but are typically between 10 and 20 minutes long. Recording entire sessions minimized the need for post-editing, and significantly reduced the number of files, but increases the need for bulk storage. Recording procedures, method, subjects and elicitation materials (referent sounds) are discussed in more detail in D2.2.2.

Table 1 lists the source and annotation files for the two female imitators, $F01$ and $F02$. As is sometimes the case with EGG measurements (for physiological reasons), the quality of the EGG signal was very poor for $F01$ and her EGG data are therefore not included in the database deliverable.

Table 2 lists the source and annotation files for the two male imitators, $M01$ and $M02$. Some technical problems arose during the recording in session 5 which made it necessary to split up the session into two subsessions. Due to this technical glitch, one imitation (referent sound *FrSt_Mi-Sol-C_sanding*) only has the audio source (recovered from a backup recording), as video and EGG were not recorded.

#	F01 Files	F02 Files
Session 01	F01_Grupp1.eaf F01_Grupp1.pfsx F01_Grupp1_Audio_16bit.wav F01_Grupp1_GOPR0075.MP4 F01_Grupp1_MVI_0018.MP4 #	F02_Grupp1.eaf F02_Grupp1.pfsx F02_Grupp1_Audio_16bit.wav F02_Grupp1_GOPR0087.MP4 F02_Grupp1_MVI_0040.MP4 F02_Grupp1_EGG_16bit_DC-corr.wav
Session 02	F01_Grupp2.eaf F01_Grupp2.pfsx F01_Grupp2_Audio_16bit.wav F01_Grupp2_GOPR0076-J.MP4 F01_Grupp2_MVI_0019-j.MP4 #	F02_Grupp2.eaf F02_Grupp2.pfsx F02_Grupp2_Audio_16bit.wav F02_Grupp2_GOPR0088.MP4 F02_Grupp2_MVI_0041.MP4 F02_Grupp2_EGG_16bit_DC-corr.wav
Session 03	F01_Grupp3.eaf F01_Grupp3.pfsx F01_Grupp3_Audio_16bit.wav F01_Grupp3_GOPR0077-J.MP4 F01_Grupp3_MVI_0021-J.MP4 #	F02_Grupp3.eaf F02_Grupp3.pfsx F02_Grupp3_Audio_16bit.wav F02_Grupp3_GOPR0089-J.MP4 F02_Grupp3_MVI_0042-J.MP4 F02_Grupp3_EGG_16bit_DC-corr.wav
Session 04	F01_Grupp4.eaf F01_Grupp4.pfsx F01_Grupp4_Audio_16bit.wav F01_Grupp4_GOPR0078-J.MP4 F01_Grupp4_MVI_0024-J.MP4 #	F02_Grupp4.eaf F02_Grupp4.pfsx F02_Grupp4_Audio_16bit.wav F02_Grupp4_GOPR0090-J.MP4 F02_Grupp4_MVI_0044-J.MP4 F02_Grupp4_EGG_16bit_DC-corr.wav
Session 05	F01_Grupp5.eaf F01_Grupp5.pfsx F01_Grupp5_Audio_16bit.wav F01_Grupp5_GOPR0079.MP4 F01_Grupp5_MVI_0026.MP4 #	F02_Grupp5.eaf F02_Grupp5.pfsx F02_Grupp5_Audio_16bit.wav F02_Grupp5_GOPR0091-J.MP4 F02_Grupp5_MVI_0046-J.MP4 F02_Grupp5_EGG_16bit_DC-corr.wav

Table 1: Data for female subjects

#	M01 Files	M02 Files
Session 01	M01_Grupp1.eaf M01_Grupp1.pfsx M01_Grupp1_Audio_16bit.wav M01_Grupp1_GOPR0039.MP4 M01_Grupp1_MVI_0003.MP4 M01_Grupp1_EGG_16bit_DC-corr.wav	M02_Grupp1.eaf M02_Grupp1.pfsx M02_Grupp1_Audio_16bit.wav M02_Grupp1_GOPR0081-J.MP4 M02_Grupp1_MVI_0028-J.MP4 M02_Grupp1_EGG_16bit_DC-corr.wav
Session 02	M01_Grupp2.eaf M01_Grupp2.pfsx M01_Grupp2_Audio_16bit.wav M01_Grupp2_GOPR0040.MP4 M01_Grupp2_MVI_0004.MP4 M01_Grupp2_EGG_16bit_DC-corr.wav	M02_Grupp2.eaf M02_Grupp2.pfsx M02_Grupp2_Audio_16bit.wav M02_Grupp2_GOPR0082-J.MP4 M02_Grupp2_MVI_0030-J.MP4 M02_Grupp2_EGG_16bit_DC-corr.wav
Session 03	M01_Grupp3.eaf M01_Grupp3.pfsx M01_Grupp3_Audio_16bit.wav M01_Grupp3_GOPR0041.MP4 M01_Grupp3_MVI_0005.MP4 M01_Grupp3_EGG_16bit_DC-corr.wav	M02_Grupp3.eaf M02_Grupp3.pfsx M02_Grupp3_Audio_16bit.wav M02_Grupp3_GOPR0083-J.MP4 M02_Grupp3_MVI_0021-J.MP4 M02_Grupp3_EGG_16bit_DC-corr.wav
Session 04	M01_Grupp4.eaf M01_Grupp4.pfsx M01_Grupp4_Audio_16bit.wav M01_Grupp4_GOPR0042.MP4 M01_Grupp4_MVI_0006.MP4 M01_Grupp4_EGG_16bit_DC-corr.wav	M02_Grupp4.eaf M02_Grupp4.pfsx M02_Grupp4_Audio_16bit.wav M02_Grupp4_GOPR0084-J.MP4 M02_Grupp4_MVI_0035-J.MP4 M02_Grupp4_EGG_16bit_DC-corr.wav
Session 05	M01_Grupp5-1.eaf M01_Grupp5-1.pfsx M01_Grupp5-1_Audio_16bit.wav M01_Grupp5-1_GOPR0079.MP4 M01_Grupp5-1_MVI_0026.MP4 M01_Grupp5-1_EGG_16bit_DC-corr.wav	M02_Grupp5.eaf M02_Grupp5.pfsx M02_Grupp5_Audio_16bit.wav M02_Grupp5_GOPR0085-J.MP4 M02_Grupp5_MVI_0037-J.MP4 M02_Grupp5_EGG_16bit_DC-corr.wav
	M01_Grupp5-2.eaf M01_Grupp5-2.pfsx M01_Grupp5-2_Audio_16bit.wav M01_Grupp5-2_GOPR0079.MP4 M01_Grupp5-2_MVI_0026.MP4 M01_Grupp5-2_EGG_16bit_DC-corr.wav	
	M01_Grupp5-sanding.eaf M01_Grupp5-sanding.pfsx M01_FrSt_Mi-Sol-C_sanding.wav	

Table 2: Data for Male subjects.

3.2 The IRCAM data set

The IRCAM data were recorded with a focus on perceptual categorization and gestural analysis. In the IRCAM data set, each imitation was recorded and stored separately, but for the purposes of articulatory annotation in D3.3.2, an entire set of 52 imitations for each subject was concatenated into a single video/audio file. Thus, only one recording of the imitation of each referent sound is included in the concatenated source file used for articulatory analysis. The articulatory annotation does not make use of those aspects of the recordings that concern gestural analysis, and thus the IRCAM data pertaining to D3.3.2 comprise only imitations for which gestures were not being specifically elicited. Recording procedures, method, subjects and elicitation materials (referent sounds) are discussed in greater detail in Sections 2 and 3 of D4.4.1.

The order in the concatenated files does not represent the order in which the imitations were recorded. Instead, the imitations for each subject have been concatenated such that the imitations of abstract referent sounds come first (Abstract01 through 12), then come interaction sounds (Interaction01 through 20) and finally machine sounds (Machine01 through 20), yielding a total of 52 imitations for each subject. At the time of writing this companion document for D3.3.2, articulatory annotation of three IRCAM subjects, 23CC, 29EB and 39CM, has been completed and 48LM is being finalized.

4 Annotation

The articulatory annotation is constituted by 8 articulatory parameters that together give a holistic description of the articulations involved. The resulting annotation can be viewed as an articulatory score that describes the contribution or action of individual articulators over time. This annotation is rich in articulatory detail to ensure that all aspects that may be significant for conveying referent sounds through imitation are covered. While the annotation system may seem complex, the separation of the different articulatory components serves to increase the accuracy of the annotation by focusing on one aspect of the articulatory flow at a time.

4.1 Annotation layers (parameters) and values

The database tool, ELAN, supports multiple layers of annotation (referred to as tiers in the ELAN documentation) aligned with both audio and video recordings. The articulatory annotation adopted in SkAT-VG makes use of 8 annotation layers, each representing a specific articulatory parameter. Two of the layers describe articulatory actions in the larynx, three layers describe the actions of the tongue (both tongue body and tongue tip), one layer is devoted to the lips, one layer controls for nasality and, finally, one layer describes the airstream mechanism (or sound initiation). In the following Section, some examples that illustrate the annotation parameters are given, but a more detailed description of these parameters and their values is given in D3.3.1.

4.2 Annotation examples

Some examples that illustrate the use of articulatory parameters and values in the database are given below. For any time point in the imitation the combination of parameter values yields a holistic articulatory description of the sound produced. In many cases, this combination yields a description that has an equivalent in phonetic transcription systems (such as the International Phonetic Alphabet, IPA). For example, in Figure 1, a male imitator ($M01$ from the KTH data set) produces an imitation of a referent sound that resembles the sound of gas squeezing through a narrow aperture. The imitator uses a fricative, [s]-like sound to imitate the hissing of the gas, and uses an occlusion at the beginning to create the sensation of a sudden onset of the sound. In the annotation it is evident that if all 8 articulation layers are combined into one, only two distinct articulatory segments emerge. The vocal folds are **abducted** throughout (Lar-VocFolds layer) and the airstream is **pulmonic egressive** (AirstreamMech), which means that air is being pushed outwards through the vocal tract. The lips are **open spread** (LipMann). The tongue makes a **dental** occlusion (Tongue-Mann) at the beginning (essentially a [t]), but then the closure is released and a dental **turbulence** is created with a **grooved** tongue (Tongue-Shape). This essentially results in a [s]-sound, which is maintained for almost two seconds. The articulatory annotation in this case follows IPA sound descriptors quite closely: a [t] is a pulmonic egressive, voiceless, dental, oral stop. The [s] is, similarly, a pulmonic egressive, voiceless, dental, grooved, oral fricative.

In Figure 2 the referent sound is reminiscent of the sound of a longish plastic pipe being struck against a solid object. Again, the imitator is $M01$ from the KTH data set. Here, one can distinguish a total of four distinct articulatory segments in the imitation. First, at the onset, an occlusion is made with the tongue tip against the upper lip (Tongue-Mann = **occlusion**; Tongue-Shape = **linguolabial**) while air is pushed up from the lungs (AirstreamMech = **pulmonic egressive**; Lar-VocFolds = **abducted**). Then the occlusion is released and, for a very brief moment (5 ms), turbulence is created as the tongue tip parts with the upper lip (the annotations are not readable at the zoom level of the screenshot). At the end of the release, the vocal folds produce a tone (Lar-VocFolds = **modal**) and the tongue assumes the shape of an [i]-like vowel (Tongue-Mann = **open**; Tongue-Shape = **F1H2**). The vowel sound is fairly brief (72 ms) and is followed by an abrupt transition into a nasal sound, IPA [ŋ], which lasts for more than a second (Nasality = **nasal**; Tongue-Mann = **occlusion**; Tongue-Constr = **velar**; the **TS#** value for Tongue-Shape indicates that this parameter is not applicable during the articulation). The total effect of the articulatory score is thus roughly equivalent to the IPA sequence [tiŋ].

In Figure 3, the same referent sound (plastic pipe) is imitated by $29EB\#$ from the IRCAM data set. In terms of articulatory segments, this imitation is more complex than the previous one, with seven segments when all annotation layers are considered together. As in the example in Figure 2, the first segment is a voiceless occlusion (Lar-VocFolds = **abducted**; Tongue-Mann = **occlusion**), in this case with the tongue tip against the teeth (Tongue-Constr = **dental**). The occlusion is then released and a short period of turbulence is created (Tongue-Mann = **turbulence**). These two segments are approximately equivalent to an IPA [t]-sound. Unlike the previous example, the occlusive is aspirated, i.e. a [h]-like sound is produced after the occlusive. This is described by a combination of values in different layers (Lar-VocFolds = **abducted**; Tongue-Mann = **open**; Lip-Mann = **open rounded**). Actually,

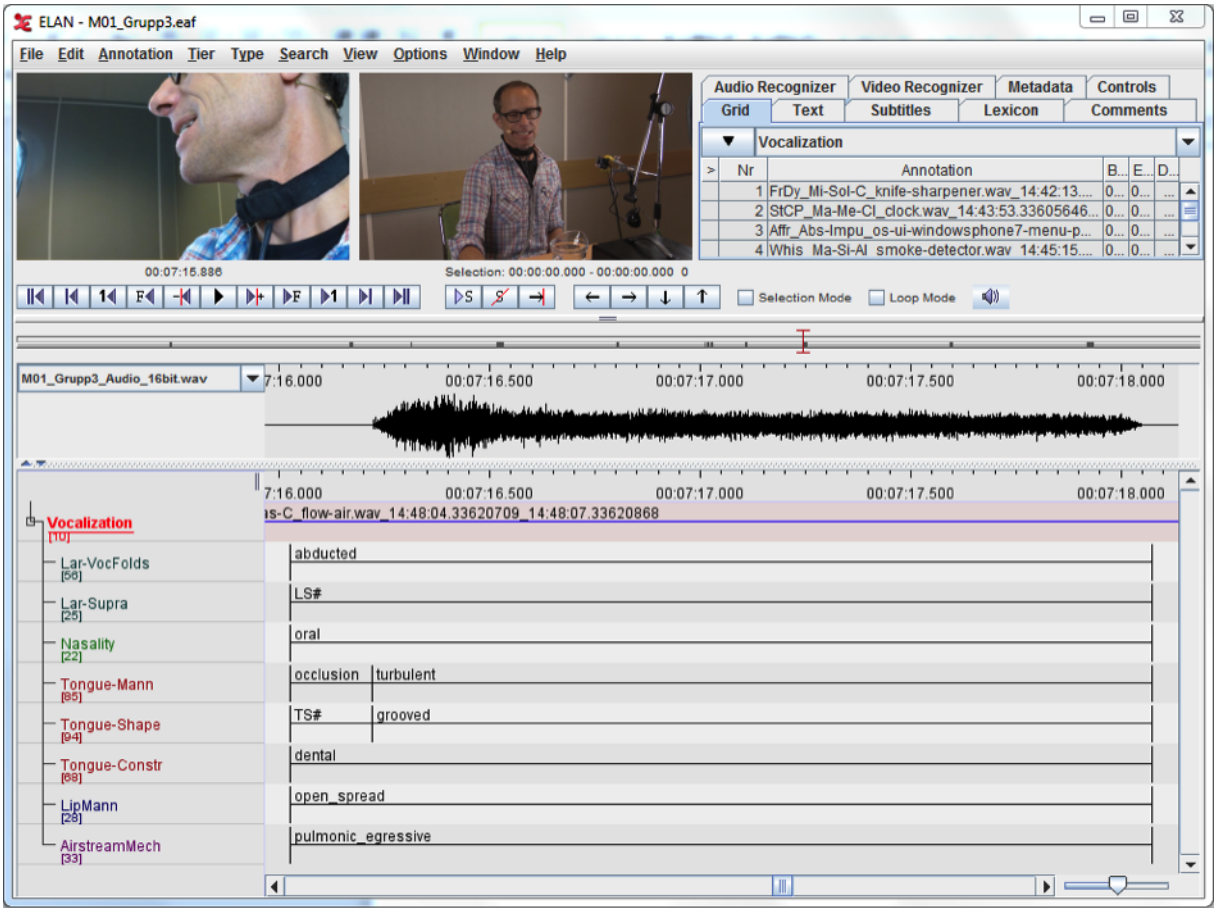

Figure 1: ELAN screen layout with annotation example 1. Side and front camera views are at top left. At the top right is the list of indexing labels. At the bottom are the eight articulatory annotation layers. The referent sound is "gas squeezing through a narrow aperture" and the imitation sounds roughly like "tssss".

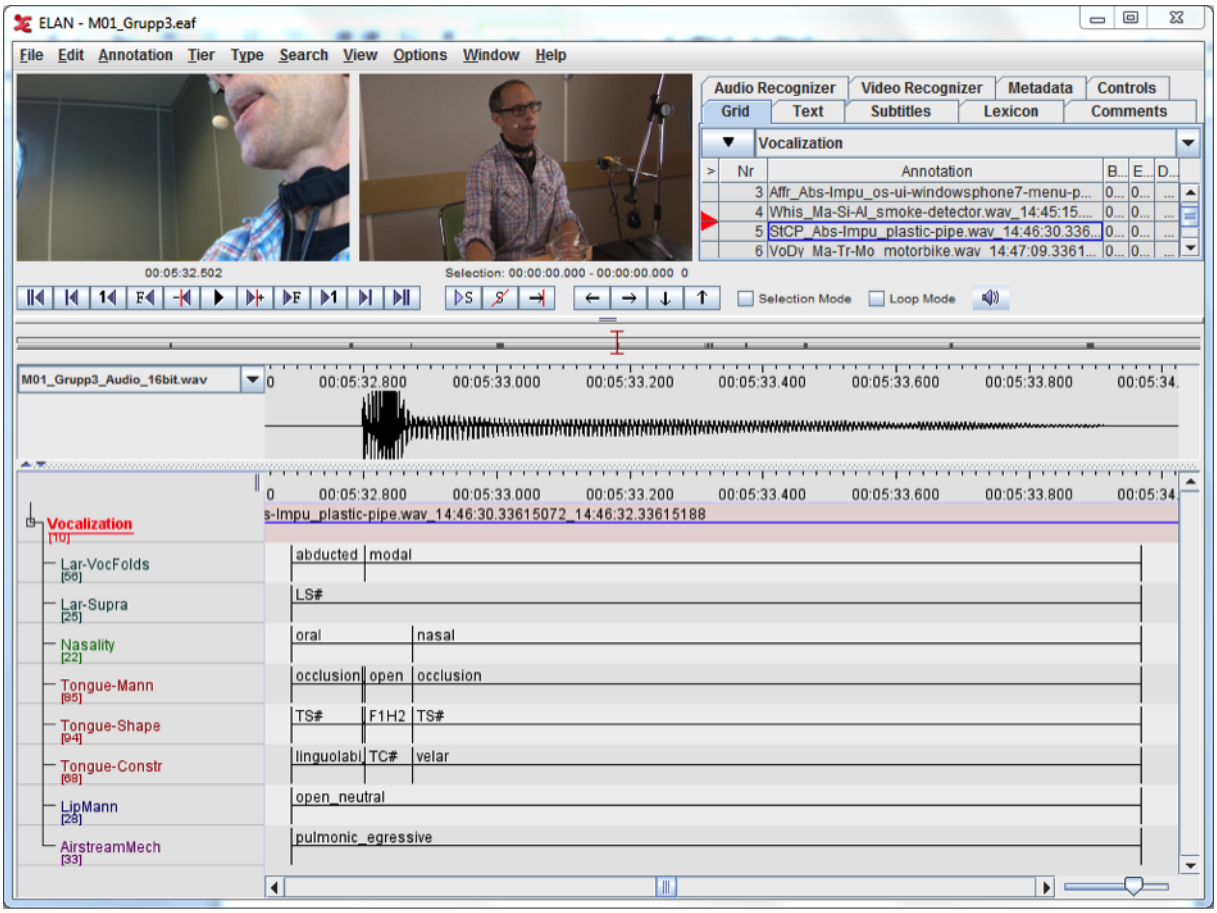

Figure 2: ELAN screen layout with annotation example 2. The referent sound is "plastic pipe being struck against a solid object" and the imitation sounds roughly like "ting".

the full articulatory annotation this sound is analysed as a voiceless [u]-vowel (Tongue-Shape = **F4H4**), which is a more precise description of the articulation than is the term aspiration.

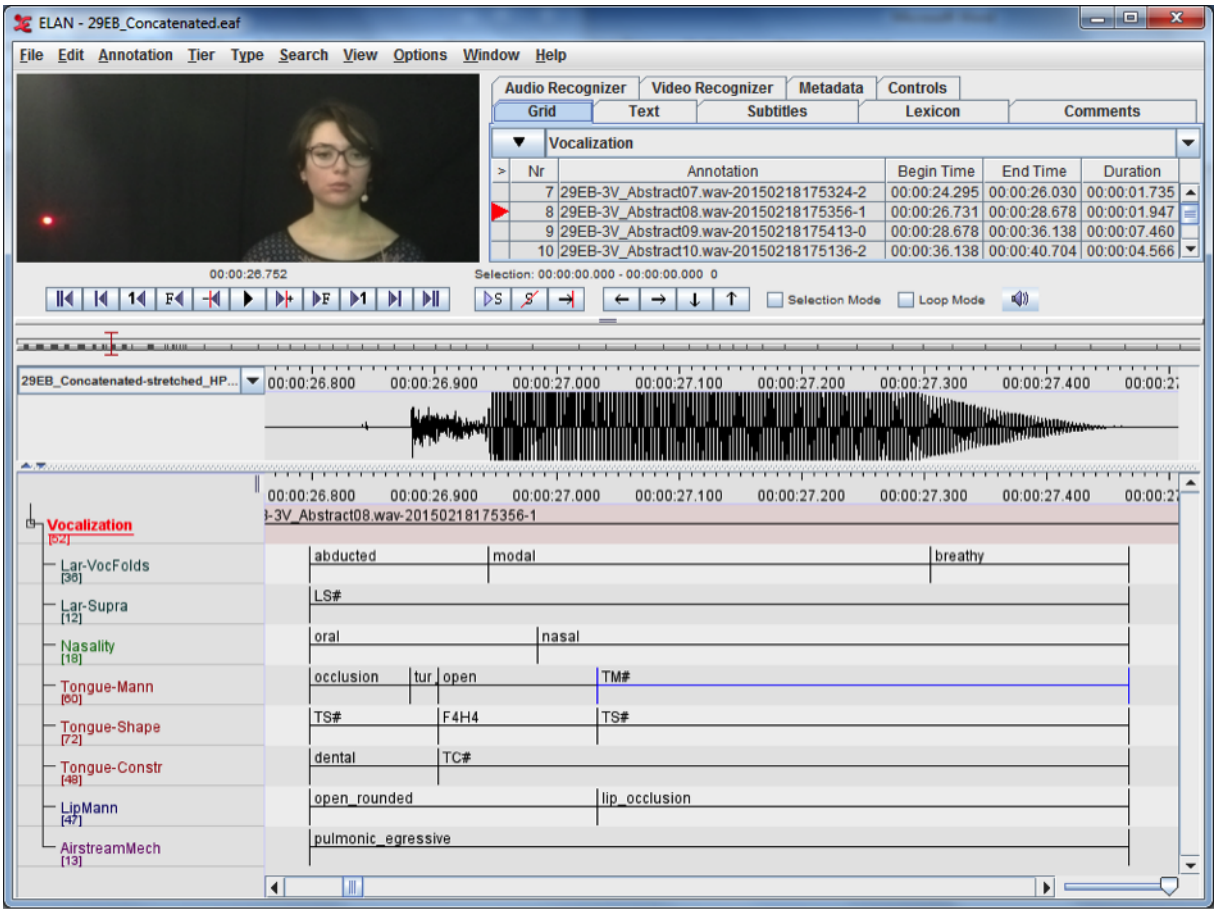

Figure 3: ELAN screen layout with annotation example 3. As in the previous example, the referent sound is "plastic pipe being struck against a solid object" and the imitation sounds roughly like "toom".

The next segment is marked by the onset of voicing (Lar-VocFolds = **modal**). This effectively changes the preceding voiceless [u]-like sound into a "normal", voiced [u]-sound. The segment that follows is marked by the onset of nasality (Nasality = **nasal**), which makes the [u] nasalized. The next segment begins as the lips are closed, creating a bilabial nasal, [m] (LipMann = **lip occlusion**). Finally, a change from modal to breathy voice marks the last segment of this imitation (Lar-VocFolds = **breathy**), which in IPA terminology would be described as a breathy voiced, bilabial nasal.

4.3 Reliability

A reannotation test was performed to assess the reliability of the annotation. (Since both annotators worked on the whole database, rechecking each other's work as it progressed, a test on the agreement between the annotators could not be performed.) The reannotation

test was performed on 40 annotations with a time lag of 6-8 weeks between the original annotation and reannotation. The test suggests that most of the articulatory parameters are robust in reannotation, even though some specific discrepancies were found between the original annotation and reannotation.

The reannotation test revealed that the annotators' judgment of the opposition between supraglottal laryngeal vibrations and supralaryngeal dorsal vibrations (approximately throat vibrations vs. trilling with the tongue) was very robust, but the opposition between ventricular vs. aryepiglottal (for supraglottal laryngeal articulations) was unreliable. Therefore, for the purposes of articulatory classification, these two articulatory values were collapsed in data extraction.

The placement of phonation onset as well as offset was highly reliable in reannotation. However, many discrepancies occurred in the classification of modal voice vs. pressed voice, so for the purposes of data extraction these values have been collapsed. The reannotation reliability of the opposition between breathy voice and modal voice was fairly robust, mainly showing discrepancies in the timing of transitions between the two.

The placement of occlusion onsets and offsets for both the Lip manner and Tongue manner tiers were highly reliable in reannotation. The reliability of the turbulent and myoelastic values was generally highly robust for these tiers as well. However, for uvular articulations the distinction between myoelastic and turbulent showed some discrepancies. The annotators therefore reviewed all cases of uvular articulations using unified criteria for their representation in the database. The use of Tongue position was generally highly robust in reannotation. The only exception to this was the boundary between the dorsal place specifications: uvular vs. velar and velar vs. pre-velar. However, changes in the place of these dorsal strictures (forward or backward) were always faithfully reproduced in reannotation.

The use of the flat and grooved values for the Tongue shape parameter was very consistent in reannotation. However, the vowel-values were often shifted in reannotation. This indicates that the distinctions made possible in the annotation system are more fine-grained than required, and that values can be collapsed to a smaller set. This smaller set will be selected if and when the articulatory classifier incorporates vowel shape in its classification. Values for Airstream were faithfully reproduced in reannotation.

The oral vs. nasal distinction (in the Nasality tier) was for the most part faithfully reproduced in reannotation. However, the precise timing of the onset of nasality (velic lowering) was not as consistent. Since nasality as such is consistently represented, we do not anticipate that the uncertainty of its exact onset will pose a problem for the articulatory classifier.

5 Data extraction

For retrieval of data from the database, the annotation system was designed with scalability in mind. First, for data retrieval, the articulatory layers can be combined in different ways to produce outcomes aimed at encompassing a specific type of articulation. For example, retrieving all examples of slow, myoelastic vibrations involves a combination of four different annotation layers (Lar-Sup, Tongue-Mann, Lip-Mann and Nasality). Second, specific layers and/or values can be collapsed or omitted in data retrieval. For example, information about tongue shape and place of constriction can be disregarded in data retrieval while information

on tongue manner is retained.

There are several ways to retrieve data in ELAN. For example, a regular expression search can be used to retrieve all instances of certain parameter values, or a sequence of values. One can also create new annotation layers that combine existing layers and then search for and extract data from the combination layers. This latter method was used to extract the data for WP5, and more details on the extraction procedure are given in D5.5.2. The search functions in ELAN are described and explained in more detail in the ELAN software documentation.