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

SkAT-VG: Sketching Audio Technologies using Vocalizations and Gestures

D3.3.1
Preliminary annotation of the database of imitations of action primitives in terms of vocal primitives

First Author	Pétur Helgason
Responsible Partner	KTH
Status-Version:	Draft-1.0
Date:	January 12, 2015
EC Distribution:	Consortium
Project Number:	618067
Project Title:	Sketching Audio Technologies using Vocalizations and Gestures

Title of Deliverable:	Preliminary annotation of the database of imitations of action primitives in terms of vocal primitives
Date of delivery to the EC:	12/01/2015

Workpackage responsible for the Deliverable	WP3
Editor(s):	Pétur Helgason
Contributor(s):	Davide A. Mauro
Reviewer(s):	Davide A. Mauro
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	06/11/2014	Import from .tex template	DAM
0.1	12/01/2015	First content draft	SteT

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 Introduction

Work Package 3 serves to document and analyze how real-world events are imitated at the phonetic level and extend our existing knowledge of the mechanisms used in the production of vocal imitations. Recordings of skilled imitators are analyzed and annotated by phoneticians, using an annotation system that describes sound production in terms of eight basic articulatory parameters. The resulting annotations, paired with the audio, then forms the input data to the machine learning developed in WP5.

The recording task has taken longer than expected, so the annotation (Task 3.2) has been started only recently. We expect to deliver an initial, sufficiently annotated set of sounds to WP5 by the end of Mo14.

2 Task 3.1 - Tool for annotation of articulatory parameters

DoW: Implementation of a data format and a graphical parameter editor for performing manual transcription of the selected articulatory parameters, from audio/video files to multitrack data files.

In discussions with WP4 and WP5, it was found that a time-continuous representation of articulators, rather than discrete attributes, would be a complication rather than an advantage when developing the machine-learning algorithms. As a matter of facts, we would have had to implement an interpreter of sorts to convert the time-continuous parameters into discrete attributes that could be used as input for machine learning. Given that discrete attributes seemed necessary as input to machine learning, the use of continuous parameters would introduce an unnecessary step in the machine learning process. Instead, the conventional scheme of discrete phonetic attributes was chosen, but supplemented, where necessary, with an optional degree of scaling for position and/or intensity. The extended annotation scheme developed for SkAT-VG is presented in Section 3 below.

2.1 Database/editor solution

In the proposal, we did not explicitly foresee the need for a database tool; however, this need quickly became evident. After taking stock of different solutions, the video/audio annotation tool ELAN was selected for WP3. ELAN is a freeware software produced by The Language Archive project at the Max Planck Institute for Psycholinguistics, Nijmegen, NL. With ELAN, the user can align several temporal records (video, audio, etc.) and make phonetic and/or linguistic annotations of the kind that SkAT-VG needs. Since the objective of annotating continuously time-varying parameters has been dropped, an editor of such parameters is no longer needed. ELAN therefore fulfills the same function as the "editor" envisaged by Task

3.1. ELAN also enables programmed queries for subsets of data to be performed, for delivery to other WP's. New versions of ELAN appear at intervals; for SkAT-VG, we are using ELAN v4.8.1 (earlier versions do not comply with our requirements).

In order to preserve the context in which imitations were performed, entire two-hour sessions were recorded with two cameras and two audio signals (voice and EGG) plus a common sync signal. This minimizes the need for post-editing, and significantly reduces the number of files, but increases the need for bulk storage. In the long recordings, the short episodes of the actual imitations are labelled semi-automatically by the experiment control software, such that it is easy for the operator to find the interesting portions. Sofia Strömbergsson worked in the summer of 2014 on implementing the annotation procedures in ELAN, and Margaret Zellers has more recently complemented and updated the procedures, as well as synchronized audio and video files for use in ELAN.

No deviations from Annex I

While some refinement will continue, this task is essentially completed.

All Objectives achieved according to Schedule

No Corrective Actions Required

3 Task 3.2 - Preliminary annotation of imitations database

DoW: Time-continuous phonological transcriptions of a number of imitations of the action primitives obtained in WP2. This will initially be done manually, using spectrograms, audio and video. These transcriptions will be in the form of continuous traces of manually estimated articulatory parameters (APs), connecting sequences of landmark points of recognizable phonetic configurations, i.e. phonemes. AP values represent the degree of activation of such phonetic sources as phonation, frication, plosives; and modifiers such as vowel resonances and stops. AP values and phonemes will form the vocal primitives of the project. In unclear cases, the imitators will be subjected to direct articulatory measurements. The vocal primitives will be stored synchronously with the original audio, accompanied by annotations of gestural primitives, and processed by IRCAM in WP4 and WP5.

For running a perceptual-feedback validation of the transcription work, it would be valuable, though not strictly necessary, to have an articulatory synthesizer that is driven to reproduce the vocal sounds interactively during the manual transcription. It is not clear that the current state-of-the-art in articulatory synthesis is good enough for this task. The existing TADA system from Haskins Laboratories will be tested as a possibility.

3.1 Annotation Scheme

The articulatory annotation system adopted for the SkAT-VG imitation database describes sound production in terms of eight articulatory parameters with (for the most part) discrete values. In ELAN, the container for the database, these parameters are represented on separate annotation tiers (see following section). In addition, information on pitch and SPL are derived from the Audio and EGG signals as time-continuous parameters, but these are not incorporated into the ELAN database, at least not in its initial form. The eight articulatory parameters reflect the basic parameters used to describe speech sounds. The parameters are: *Vocal folds, Supralaryngeal phonation, Nasality, Tongue manner, Tongue shape, Tongue constriction, Lip manner and Airstream.*

The *Vocal folds* and *Supralaryngeal phonation* parameters describe the configuration of the larynx. The *Vocal folds* parameter describes the state of the vocal folds, for which the most usually occurring values are **modal** phonation, **falsetto** and **open** (which allows for an unrestricted airstream). Supraglottal phonation relates to the use of **aryepiglottal** and **ventricular** vibration, which are frequently used for imitating sounds with periodic vibration frequencies lower than the imitator can produce using the vocal folds. A full list of the possible annotation values for the two laryngeal parameters is given in Table 1.

Tier Group	Articulator Tiers	Tier Values
Larynx	Vocal Folds	modal
		modal
		falsetto
		pressed
		creaky
		breathy
		whisper
		closed
		open
	Supraglottal phonation	aryepiglottal
		ventricular

Table 1: Laryngeal annotation values

The *Nasality* parameter (Table 2) describes velic state, i.e., whether the velum is raised, thus preventing nasal resonance, or whether it is lowered, which induces nasal resonance. Thus, the parameter only has two values, **oral** and **nasal**. One of the acoustic impacts of nasality is

to broaden the resonance frequency peaks of the vocal tract so imitators can use nasality to "flatten" the frequency response of the vocal tract, which is particularly useful when imitating engine sounds.

Tier Group	Articulator Tiers	Tier Values
Nasality	Nasality	oral nasal

Table 2: Nasality annotation values

There are three parameters relating to tongue configuration: *Tongue manner*, *Tongue shape* and *Tongue constriction*. The possible values for these parameters are given in Table 3 *Tongue manner* describes the type of constriction used when the tongue is involved in the production of an imitation. The most frequently occurring values for *Tongue manner* are **turbulent** (which covers fricative-like sounds), **myoelastic** (which covers trilling sounds), **occlusion** (which covers stop-like sounds) and **open** (which is used when no constriction is in place, e.g. in the production of vowel-like sounds). The values **flap** (implying a backwards or forwards **flap** of the tongue against the roof of the mouth) and **turbulent**-whistling (which can be equated with whistling sibilants in speech) occur far less frequently. *Tongue shape* and *Tongue constriction* are by far the most diversified in terms of the possible values they can take. *Tongue shape* differentiates between three different tongue shapes associated with consonant-like sounds. The value **flat** (which can be described as a neutral tongue shape) is the most common. The value **grooved** (implying that the airstream is channelled using the anterior part of the tongue) is used for sibilant sounds. The **lateral** value is used for sounds which are produced analogous to lateral speech sounds (i.e., with the tongue forming **lateral** channels for sound propagation and/or airstream). Lastly, vowel-like sound production is described with a set of values under the *Tongue shape* parameter. To cover the range of possible values in the continuous domain that is vowel space, an approach was adopted that divides both vowel frontness and vowel height into five discrete steps: F1 to F5 for frontness and H1 to H5 for height. This results in 25 parameter values, ranging from **F1H1** (an high, front vowel-like articulation) to **F5H5** (a low, back articulation). *Tongue constriction* describes the placement of a constriction of the tongue in the vocal tract. This largely corresponds to the place divisions used for describing speech sounds, for example **uvular**, **velar** and **alveolar**. Apical speech sounds (made using the very tip of the tongue) are given a value separate from their laminal counterparts (which are made using the most anterior part of the tongue, the tongue blade). The apical parameter values are designated as "apico-" in the descriptor, e.g. **apico-alveolar**, but corresponding laminal values are not preceded by a "lamino-" designation (and thus, e.g., the value **alveolar** implies a lamino-alveolar articulation).

One should note that both the *Tongue shape* and *Tongue constriction* parameters have the value transition. This value implies a gradual transition from a preceding state to a following state. For example, an articulation that involves a gradual shift from one vowel-like configuration (e.g. **F3H4**) to another (e.g. **F1H1**), which would result in a diphthongal, [ai]-like sound, would be described by the *Tongue shape* sequence **/ F3H4 / transition / F1H1/** in our annotation scheme. Likewise, gradual shifts in friction noise can be described using the transition value. For example, the gradual shift in **turbulent**, fricative noise in an articulation shifting from a **velar** constriction into a palatal constriction can be described by the *Tongue*

Tier Group	Articulator Tiers	Tier Values
Tongue	Tongue manner	turbulent
		myoelastic
		flap
		occlusion
		turbulent whistling
		open
	Tongue shape	flat
		grooved
		lateral
		transition
		VOWELS, see description in text
	Tongue constriction	dental
		alveolar
		post-alveolar
		palatal
		pre-velar
		velar
		uvular
		pharyngeal
		linguolabial
		apico-dental
		apico-alveolar
		apico-post-alveolar
		apico-palatal
		apico-pre-velar
		labiodental
		transition

Table 3: Tongue annotation values

constriction sequence **/ velar / transition / palatal/**. This aspect of the annotation scheme thus goes some way towards accommodating the time-continuous parameters that were initially proposed.

The *Lip manner* parameter describes the articulatory action of the lips. The lips can be an active articulator that results in both **turbulent** and **myoelastic** (trilling or vibrating) sounds. The values **turbulent spread** and **turbulent rounded** describe fricative-like articulations, primarily intended to cover articulations that correspond roughly to voiceless bilabial and voiceless labiovelar fricative speech sounds, respectively. The values **myoelastic tense** and **myoelastic lax** relate to productions that are acoustically quite different. The **myoelastic lax** value describes lip trilling (a bilabial trill in the IPA system), which is easy enough to produce but still uncommon as a speech sound. Imitators use bilabial trilling quite often for mimicking engine and animal sounds. The **myoelastic** tense value describes a lip vibration (faster than trilling) similar to that used for playing many brass instruments (a technique referred to as embouchure). This occurs fairly infrequently in our imitation data. The **occlusion** value implies that the lips are completely closed. This contrasts with the four values which describe lip aperture, whose descriptors are fairly self-explanatory: **open rounded**, **open half rounded**, **open neutral**, and **open spread**. The value **whistling** describes labial whistling, which is usually employed when imitating tweeting birds but which is otherwise fairly uncommon. The **transition** value is used primarily to describe articulations in which rounding is gradually changed, for example from **open rounded** to **open neutral**.

Tier Group	**Articulator Tiers**	**Tier Values**
Lips	Lip manner	turbulent spread turbulent rounded myoelastic tense myoelastic lax occlusion open rounded open half rounded open neutral open spread whistling transition

Table 4: Lips annotation values

Lastly, the *Airstream* parameter describes the airstream mechanism used for sound initiation, which can be seen as the most basic aspect of the articulation. The most commonly occurring value for *Airstream* by far is **pulmonic egressive**, which implies that the lungs push out air through the vocal tract. This is also the dominant sound initiation mechanism in any spoken language. The **pulmonic ingressive** value (air is drawn in) is far less common, but still occurs in varying degrees for different imitators. The **glottalic egressive** and **glottalic ingressive** values are very infrequent, but were observed to occur in the SkAT-VG exploratory imitation database. Likewise, the **velaric egressive** value is very infrequently used, but was observed in the exploratory data. The **velaric ingressive** value, however, which corresponds to click sounds in speech, occurs quite frequently, for example in the imitation of mechanical

devices such as clocks. Finally, there are three values describing **percussive** articulations, i.e. articulations that do not actually employ an airstream. Three values that were observed in the exploratory data are included here: **percussive bidental** (the clashing or gnashing of the teeth), **percussive laminodental** (slapping the tongue against the teeth) and **percussive** sublaminal (slapping the tongue against the floor of the mouth).

Tier Group	Articulator Tiers	Tier Values
Airstream	Airstream mechanism	pulmonic egressive pulmonic ingressive glottalic egressive glottalic ingressive velaric egressive velaric ingressive percussive bidental percussive laminodental percussive sublaminal

Table 5: Airstream annotation values

The articulatory annotation system was designed with a view to scalability, meaning that the potential complexity that arises from the combinatory possibilities of the values for different parameters can be drastically reduced by collapsing values for individual parameters and/or omitting whole parameters. For example, form the point of view of machine learning, the division of the vowel space into 25 areas might prove unfruitful and collapsing certain vowel categories might improve results. Similarly, for machine learning, the *Airstream* parameter might be adequately represented by just two values (**pulmonic** vs. **velaric**). Likewise, the values for *Tongue constriction* might be collapsed to form only four contrasts (instead of the 15 values used in the annotation). Such simplifications are easy to achieve using parametrized values such as those used here. In an IPA-like annotation, which may appear simpler because it requires only one annotation tier, each unit represents a constellation of articulatory parameters and lacks direct scalability. Collapsing articulatory aspects of such an annotation would have required a complex translator that decomposed each annotation value into its constituent articulatory components.

3.2 Annotation examples

In this section we give examples of the articulatory annotation in the imitations database. The sound files corresponding to these examples can be found as the accompanying material for this deliverable. The first example (Figure 1) is a male imitator (M02) producing an imitation of the sound of gas squeezing through a narrow aperture (such as the sound from a spray can or from venting gas from a gas canister). The imitator's strategy is essentially to mimic the mechanical action that gave rise to the referent sound, i.e. squeezing air through a narrow channel. He does this by producing a fricative sound that resembles a sibilant speech sound, the production of which requires directing an airstream through a narrow channel. However, contrary to a normal speech sound, the airstream is ingressive, which has the effect

of broadening the spectral peaks, making the sound less [s]-like than it would be on an egressive air stream.

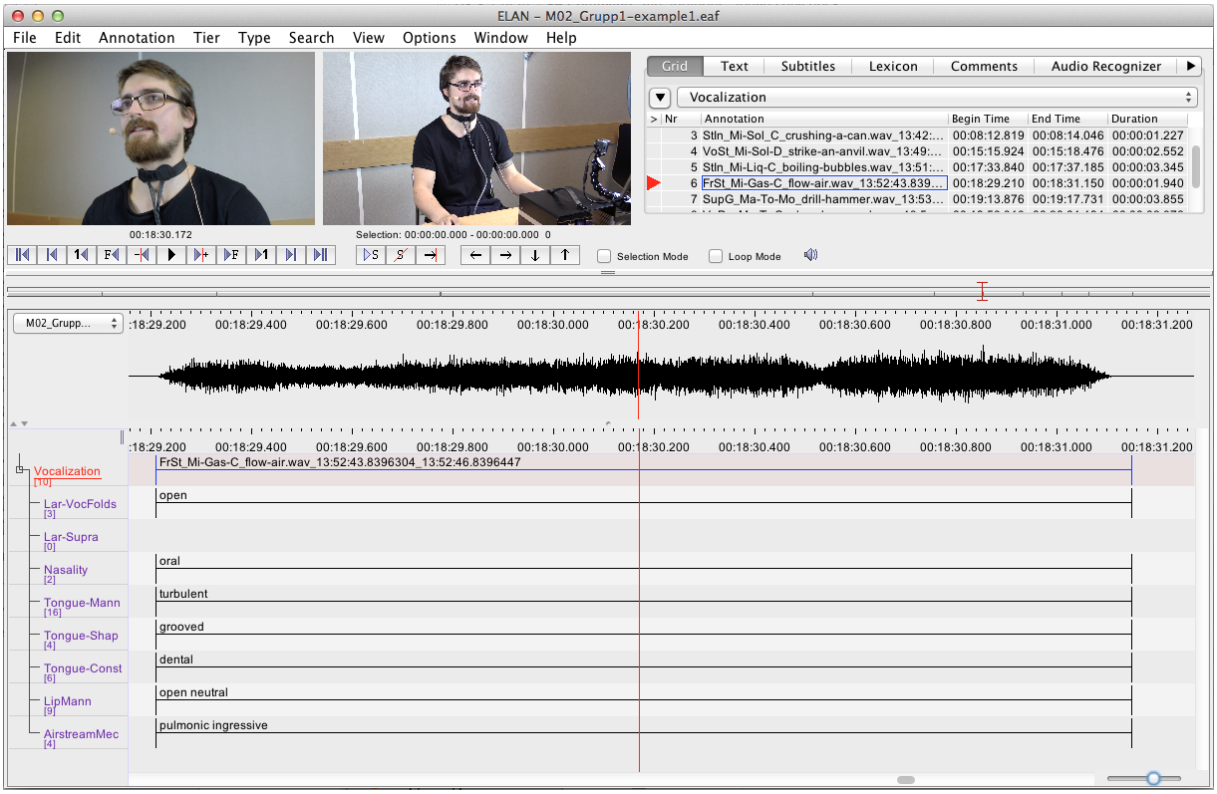

Figure 1: ELAN screen layout with annotation Example 1, "spray can". Side and front camera views are at top left. At the top right is the list of indexing labels generated by the experimental control software (WP2). At the bottom are the eight annotation tiers. In this example, the imitator's articulation remained unchanged throughout the imitation. From the video, it is evident that the imitator produces the sound by drawing in rather than expelling air.

The articulatory notation reflects the ingressive airstream in the Airstream annotation tier (*AirstreamMec*, the bottom tier in Figure 1). The Vocal Folds tier (*Lar-VocFolds*) indicates an open configuration for the larynx allowing air to flow through freely. The fricative nature of the sound is indicated on the Tongue manner tier (*Tongue-Mann*), which indicates a constriction that gives rise to a turbulent sound source. The placement (and thus the approximate acoustic character) of the stricture is indicated on the Tongue constriction tier (*Tongue-Const*). The lips remain neutral during this articulation, indicated on the Lips tier (*LipMann*), and the sound has no nasality, as indicated on the Nasality tier. Finally, there is no involvement of the supraglottal phonatory organs (*Lar-Supra* tier). The combination of a pulmonic ingressive airstream, an open larynx and a dental constriction is that air gets sucked in creating turbulence in the area between the teeth and the tongue blade.

The second example (Figure 2) is a female imitator (F01) producing an imitation of an "X-ray buzzer", a harsh, periodic sound with constant pitch. The imitator uses the vocal folds to capture the periodicity of the "buzzer", but to achieve the harsh-sounding effect of the

referent sound she presses the vocal folds together harder than one would in normal speech (*Lar-VocFolds* tier) and also introduces a vibration in the aryepiglottal folds (*Lar-Supra* tier). There is nasality throughout the entire imitation (*Nasality* tier) and, in fact, it starts with an articulation that sounds like a dental nasal consonant, which then changes to a nasalized, vocalic articulation (as indicated in the three *Tongue* tiers). As indicated in the *Tongue-Shape* tier, the vocalic articulation is central both in terms of height and frontness (i.e., a schwa-like quality). The lips are somewhat protruded, as indicated on the *LipMann* tier.

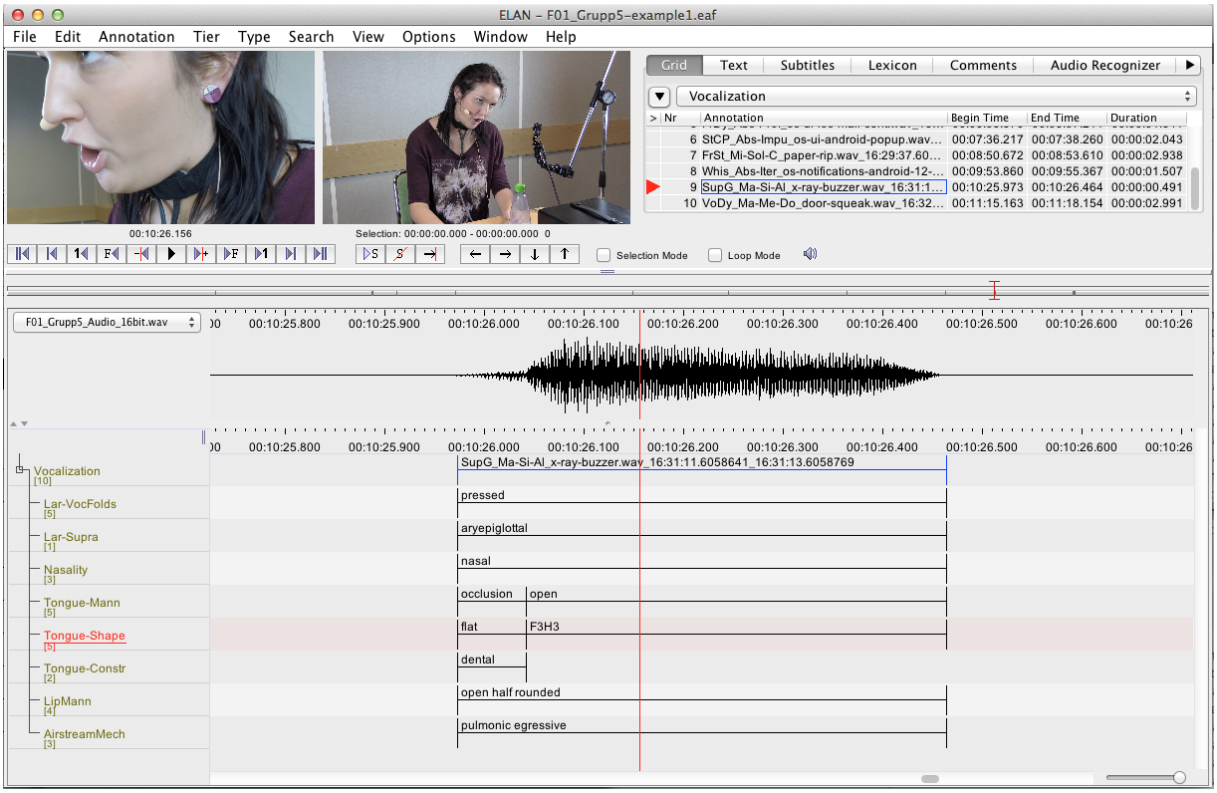

Figure 2: Annotation Example 2: "X-ray buzzer."

The third example (Figure 3) is the same female imitator producing an imitation of the ringing of an alarm clock bell. For the most part, the imitation consists of an alveolar trilling sound (IPA [r]), which is indicated in the *Tongue* tiers as myoelastic and alveolar. Towards the end, the myoelastic constriction is lifted and a fairly central vowel sound is produced.

Lastly, the fourth example (Figure 4) is the male imitator (M02) producing an imitation of the sound of crushing an aluminium can. This is a quite complex sound that can be seen as a series of impacts, each of which may be seen to constitute a vocal primitive. The speaker produces a series of sounds using an ingressive airstream, mixing click sounds (velaric ingressive) produced at various places of constriction with lateral, ingressive turbulence, all the while varying lip position. The effect is a fast-changing and complex acoustic image that attempts to replicate the complex series of acoustic events that make up the referent sound.

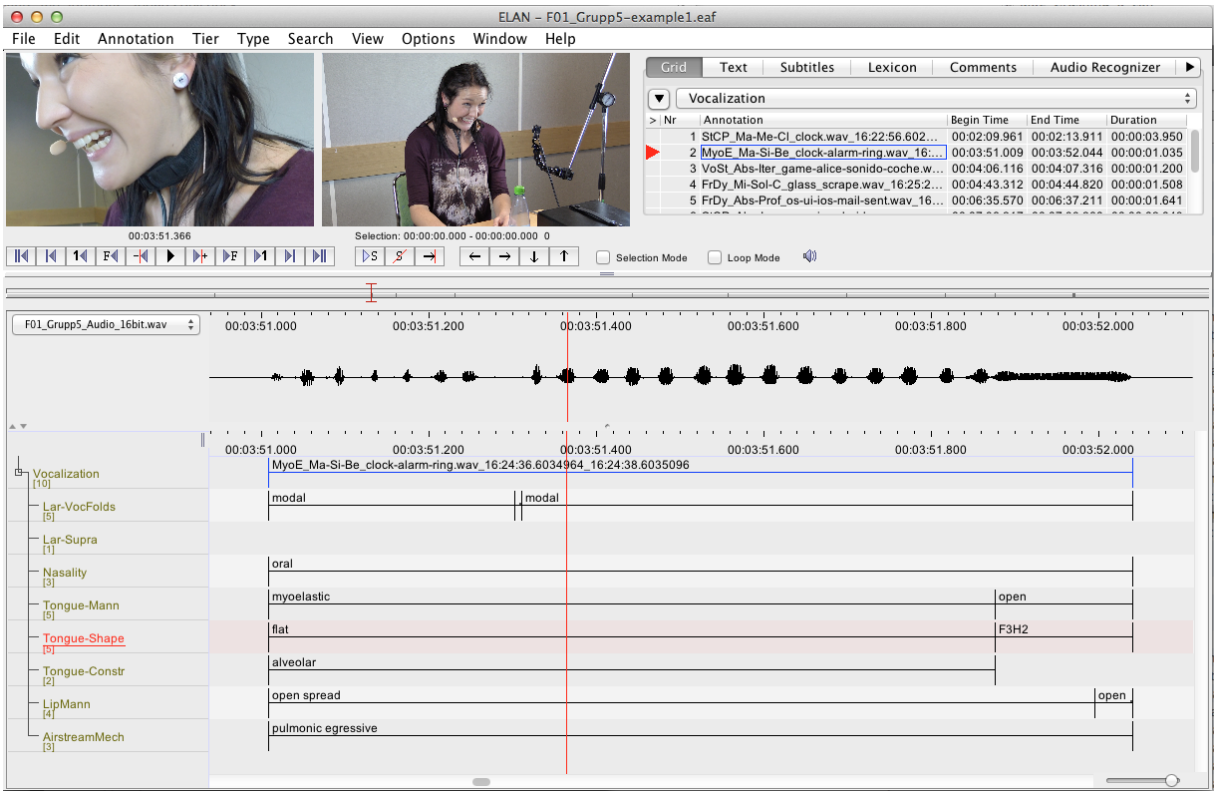

Figure 3: Annotation example: "alarm clock ringing."

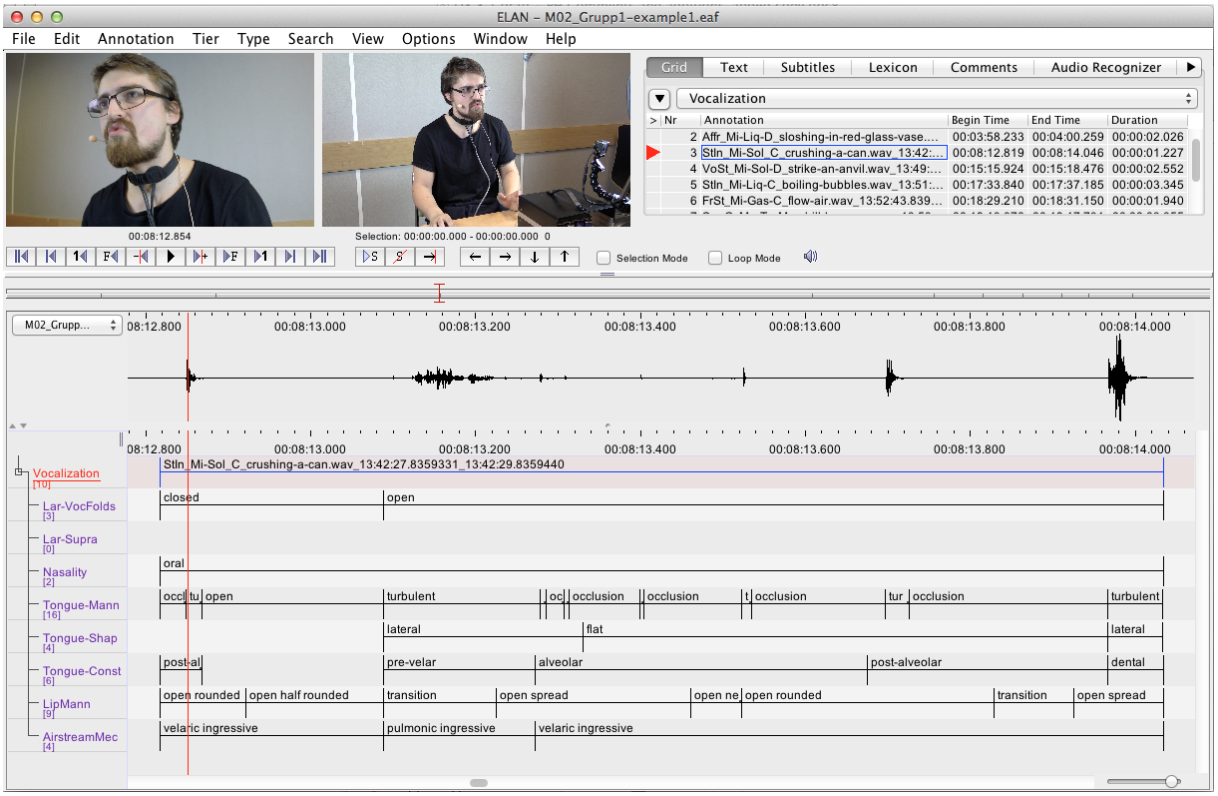

Figure 4: Complex annotation Example: "crushing a can."

3.3 WP3 Action Points for Mo13-24

- Devise mechanism for project-wide Internet access to the imitations database. If issues of bandwidth and/or security prevent this, devise procedure for dissemination of versioned copies.
- Continue annotation work as recordings of the required number of imitators become available from WP2 (Mo13-14).
- Annotate recordings from Task 4.2 as they become available.
- Deliver annotated recording excerpts to WP5 as they become available.