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

SkAT-VG: Sketching Audio Technologies using Vocalizations and Gestures

D3.3.3
Linguistic aliases for non-vocal sounds

First Author	Pétur Helgason, Christine Ericsdotter, and Sten Ternström
Responsible Partner	KTH
Status-Version:	Final-0.1
Date:	January 5, 2017
EC Distribution:	Consortium
Project Number:	618067
Project Title:	Sketching Audio Technologies using Vocalizations and Gestures

Title of Deliverable:	Linguistic aliases for non-vocal sounds
Date of delivery to the EC:	31/12/2016

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

Abstract	The current deliverable presents the results of tasks T3.3
Keyword List:	linguistic aliases

Document Revision History

Version	Date	Description	Author
First draft	22/12/2016	Import from .tex template	ROC
Final	28/12/2016	Version to be submitted	Pétur

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

As in previous work packages, we adhere to a terminology that differentiates between different actors and actions in the communication of sound images: referent sounds are the sounds being imitated, the imitator is the person who performs an imitation (a vocal and/or gestural rendering of the referent sound), and the receiver is the one who perceives the imitation and tries to interpret it.

In Task 3.3 we examined whether the SkAT-VG system would benefit from using word-like aliases (onomatopoeia) as an input channel to the system. To maximize the chances of obtaining data with onomatopoeia, and to examine the effect of different types of receiver and referent sounds on the imitations, we constructed an experiment in the form of a game between an imitator and a receiver, in which the imitator described a set of referent sounds to the receiver. There were three rounds played, each with a different receiver. In the first round the receiver was a person who shared a native language with the imitator; in the second round, the receiver was a person who had no shared language with the imitator; and in the third round the receiver was a (Wizard of Oz) computer that only processed sound and not language. The imitations were evaluated on a scale of wild to tame imitation using a set of phonetic criteria. The results indicate that an imitator and receiver communicating using a shared native language yield far more instances of tame (linguistically grounded) imitations than when imitator and receiver (human or computer) do not share a language. Given these results, we suggest that the use of linguistic aliases would not be useful as an input channel to the SkAT-VG system (a computer without automatic speech recognition). Further, it was evident that the nature of the referent sound correlated with the propensity for tame output: the more the morphological structure (with respect to duration and intensity) of referent sounds resembled the structure of syllables in speech, the more prone they were to being rendered in a tame fashion by imitators. We conclude that the morphology of the referent sound is a better predictor of tameness than the putative tendency for imitators to render vocally inaccessible sounds using tame imitations. Lastly, for the shared language condition, we observed a decidedly bimodal distribution in tameness scores, with imitations being produced either as quite wild or quite tame rather than being in an intermediate stage between wild and tame. Our interpretation is that the imitators choose one of two articulatory strategies, which we refer to as sound mode (for wild imitations) and language mode (for tame imitations).

2 Introduction

2.1 Background to Task 3.3

In Task 3.3, we have considered whether sounds that are not easily imitated through vocalization might be specified using word-like aliases or semi-symbolic sounds and whether such onomatopoetic expressions, in the form of a generalized (language neutral) vocabulary, would be a viable input alternative for users of the SkAT-VG system. Task 3.3 has also examined how referent sounds differ in terms of how amenable they are to being rendered in an onomatopoeic form and seeks to explain these differences.

Onomatopoeia, as we define it here, comprises utterances that are wholly or partially grounded in the phonological/phonetic system of a given language. To be considered ono-

matopoeic, an utterance should adhere to the basic characteristics of syllable structure, segmental contents and prosodic shape of the language in question. The more an utterance deviates from these aspects of linguistic well-formedness, the further it is pushed into the realm of imitation or mimicry ([And98], 141ff).

In linguistics, onomatopoeia is one of many concepts that relate to sound imitation and the relationship between sound and meaning in languages. Onomatopoeia is part of a broader category of lexical items called ideophones. In many languages, ideophones form a lexical class in its own right, with different ideophone items evoking different sensory events ([Dok54] on Bantu languages, [Ham98] on Japanese, and [Aus94] on Finnish, to name a few examples). Some ideophones are clearly onomatopoeic, i.e. their phonetic form is sound-mimetic, but such a connection is not mandatory at all.

Another relevant linguistic concept is the phonestheme, which denotes the tendency in languages for association between meaning and individual sounds or clusters of sounds. For example, a word-initial sl- in English has a fuzzy association with thick liquids and pejorative traits. Some sound-meaning associations are fairly well established cross-linguistically, such as the tendency for the concept of small to be rendered with high, front vowels [Dif94]. A more recent, broader corpus study across approximately two-thirds of the world's languages has shown that such meaning association may be further reaching and more universal than has been previously suspected [BWH$^+$16], indicating that associations of this type can be grounded in factors common to humans as a species.

In light of these facts, the idea of trying to establish a universal phonetic basis for expressing different types of sounds is worth consideration. One should keep in mind, though, that while sound symbolism is evident in languages its power is actually limited, since only the strongest associations reported by Blasi et al. [BWH$^+$16] would be useful from a cross-linguistic perspective. Constructing an sequence of sounds as an alias for a given type of sound when the association is weaker would be unlikely to reflect the actual phonological encoding in a given language. Also, the associations reported only cover a very small portion of the total semantic space.

Our immediate aim, though, is not to investigate correspondences between sound and meaning, as such. Instead, our primary focus is on sound-to-phoneme correspondences (onomatopoeia) and the relationship of imitator sound output to the nature of the receiver and of the referent sound.

From our investigations in SkAT-VG prior to Task 3.3, the general observation could be made that onomatopoetic utterances were quite rare in our data and that those instances that did occur might equally well be interpreted as imitative rather than linguistically encoded. Looking to the IRCAM and KTH databases collected within the project this is not surprising, since the informants were given the task of imitating (or sketching) referent sounds. The tasks were not embedded in a linguistic context, i.e. the imitators were not speaking, describing the sounds verbally or otherwise engaged in any linguistic exchange. The instructions they were given were to listen to the referent sounds and imitate them.

To elicit data that was richer in onomatopoeic content, WP3 collaborated with WP7 on conducting workshops with students of sound design in Copenhagen. The students were recorded during a collaborative sound design exercise, during which they, through discussion and interaction, developed different sounds appropriate for their task. We expected that these design activities, as well as playing a round of the *Imitation Game*, would yield more

occurrences of tame imitations, given the linguistic context in which these exercises embed the imitative effort. Disappointingly, though, as we analysed these data we found that examples of tame imitations were quite scarce even in this material.

Given the results of the experiment described below, i.e. that tame imitations occur mostly when the receiver shares a native language with the imitator, the scarcity of tame imitations in the workshop data is perhaps understandable. The participants in the workshop were of different nationalities, using English as a *lingua franca*. While quite fluent in English, the participants would still be likely to struggle to access the vocabulary and means of expression (including onomatopoeia) that pertain to the description of sound in English, which they might do much more easily speaking to someone in their native language.

Thus, so far in the project, none of the data gathered had turned up much in the way of onomatopoeic expressions. We posed three questions:

- Can imitators be induced to produce onomatopoeic expressions at all when communicating sound images to others?
- Is there a categorical boundary between onomatopoeia and imitation/mimicry, or is there a gradual transition from the one to the other?
- If we could induce imitators to produce onomatopoeic expressions, would different types of referent sound be treated in different ways?

To answer the first question, we devised an experiment that would likely maximize the use of onomatopoeic expressions. This involved embedding the description of referent sounds in a linguistic context of verbally describing referent sounds, making sure that the linguistic context was native to the imitator. Also, we removed the requirement of imitating the sounds, but instead gave the instruction to describe the sound using any expressive means available (gestures, verbal descriptions, etc.). As it turned out, the imitators would relatively seldom fail to produce some kind of imitation of the referent sound, despite not being required to do so.

The second question relates to the articulation strategies chosen by imitators when imitating a sound. The onomatopoeic utterances used in language and the imitative vocalizations observed in the SkAT-VG articulatory data may indicate two different ways in which humans approach sound production.

Onomatopoeia in languages makes use of expressions whose phonetic appearance can be derived from similarities with real world sounds. For any given language, such onomatopoetic expressions are shaped and constrained by the language's phonological rules and phonetic expression and are therefore inevitably language-specific. Speakers of English, Estonian and Cantonese, for example, perceive the sound made by a duck in the same way but its onomatopoetic forms are *quack quack*, *prääk prääk*, and *gua gua* respectively.

By contrast, in the articulatory SkAT-VG imitation database (comprising four Swedish and four French imitators), there is little to suggest that the imitators are constrained in their imitations by language specific factors. The imitators have access to the articulatory repertoire of their native language as well as knowledge of a number of onomatopoetic forms. Still, given a task of sketching sound through imitation, they seem free to explore their sound space and their vocalizations do not appear to be constrained by native linguistic structures. Instead, the

imitators readily use sound production mechanisms that, in linguistic terms, are non-native to them and they rarely produce imitations that resemble onomatopoetic words.

This points to an apparent split between linguistic and non-linguistic sound production. On the one hand we have a "language mode" of production, which is constrained by native phonology and phonetics. On the other we have a "sound mode" of production, which is not constrained by native language phonology or phonetics, but is instead subject only to the general articulatory constraints imposed by the human vocal apparatus. Acquiring data covering the range from onomatopoeic (tame) to imitative (wild) could help to shed light on whether the transition between the two modes of production was gradual or categorical.

The third question concerns the nature of the referent sounds being imitated. If imitators can be induced to produce onomatopoeic utterances, will we observe differences in the propensity for onomatopoeia for different referent sounds. In particular, we wished to examine the hypothesis that the more the morphological structure (with respect to duration and intensity) of the referent sound resembled the structure of syllables in speech, the more prone they would be to be rendered through onomatopoeia. Thus, for the experiment, we selected a range of referent sounds with morphological structures ranging from speech-like to non-speech.

The remainder of this deliverable describes the experimental setup and the results from the experiment.

3 Method

3.1 Collaborative game

To elicit a wide range of imitation types, a collaborative sound guessing game was designed for the experiment. The rules of the game were as follows: The imitator and receiver stand on either side of a sound proof window in a recording studio. They each have a tablet device with an app developed for this experiment, with buttons for playback and feedback (Figure 1 and next section). The task of the imitator is to listen to a referent sound through headphones, and to describe it to the receiver, using any communicative resources available. The receiver's tablet has 10 sounds to choose from. The receiver has to listen to these sounds and guess which sound the imitator heard based on the imitator's description. The receiver indicates his/her choice by playing the chosen sound to the imitator. If the receiver's guess is correct, the players get one point in the game, and can move on to the next referent sound. If the guess is wrong, they can have one more try at that referent sound. If the receiver can't guess the sound the second time either, the players get 0 points for that referent sound, and the next referent sound is presented.

The imitator, a native speaker of Swedish, plays three rounds of the game with three different receivers. In the first round, the receiver is a native speaker of Swedish. In the second round, the imitator is made to believe that the receiver does not understand Swedish, or any other Indo-European language. In the third round, the receiver is a (Wizard-of-Oz) computer. In this way, the number and type of communicative resources available to the imitator change between the rounds of the game. The change from shared language to non-shared language and from human to computer is designed to set the communication in different degrees of linguistic embedding.

Figure 1: Experimental setup

The imitator is not aware that in all test conditions the receiver is an experimenter who knows where the matching sound is placed on the tablet. This gives the experimenters the power to control the feedback. This allows them to encourage the imitator to be creative in the sound descriptions and secure that an adequate amount of material is being produced.

3.2 Referent sounds

For each of the three rounds in the game the imitator described 10 referent sounds to the receiver. The sounds are listed in Table 1. Slightly different version of the referent sounds were used for each round of the game. However, the basic characteristics of the referents were very similar, as listed in Table 1 and depicted in Figure 3. Each of the sounds Bell, Boom, Flow and Mixer can be characterised as a single, prolonged event (cf. associated durations in Table 1). One sound, Click, is a combination of two discrete sound events with no pause in between, and the duration cited in Table 1 is the mean duration for the two events. The sounds Drip and Tick are composed of several events separated by pauses, and the durations cited in Table 1 are the mean durations for these events. The sounds Alarm, Saw and Wipers are composed of several cycles of near identical events, which are not separated by pauses. The mean duration of the cycles within each sound is given in Table 1. Finally, Table 1 indicates whether or not the referent sound contains a periodic source.

Figure 2: View of receiver's tablet

Table 1: List of referent sounds and their basic characteristics

Referent sound	Total duration (ms)	Number of cycles/events	Duration (ms) of cycle/event	Periodicity
Alarm	1600	8	200	yes
Bell	2100	1	2100	yes
Boom	1500	1	1500	no
Click	280	2	140	no
Drip	3700	7	70	no
Flow	3900	1	3900	no
Mixer	3800	1	3800	yes
Saw	3600	6	600	no
Tick	3100	10	70	no
Wipers	5300	6	890	yes

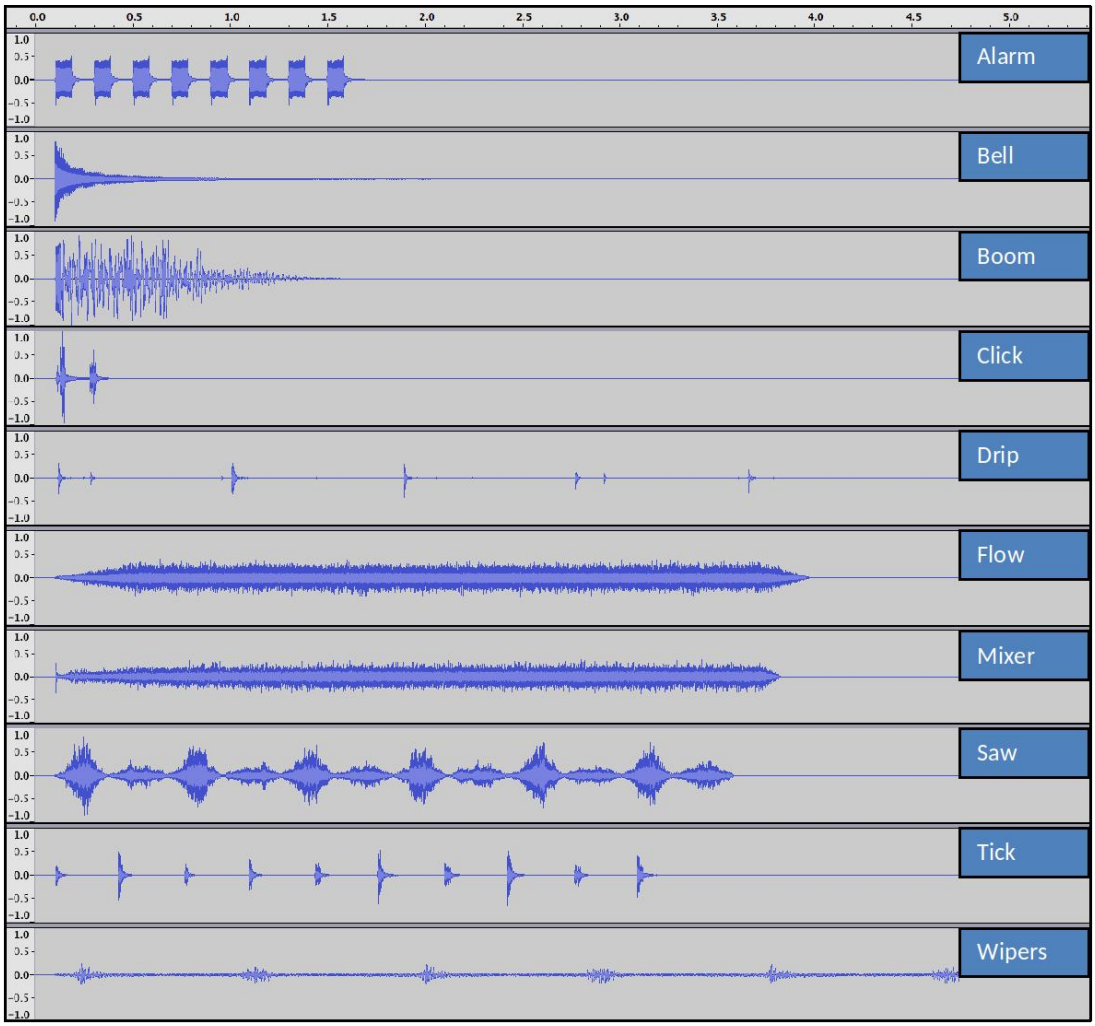

Figure 3: The waveform of each type of referent sound used for the experiment. Slightly different versions of each sound (varying in timbral quality) were used during each of the three rounds but they were equivalent in terms of overall shape (duration and intensity).

3.3 Game software

The software for the collaborative game consisted of three parts: an imitator's app (see Figure 4 left), a receiver's app (see Figure 4 right) and a server logging all events in these applications. For each referent sound in the imitator's app, 10 different sounds were loaded to the playback buttons in the receiver's app (marked with numbers from 1 to 10). The imitator could play back the referent sound as many times as s/he wanted by tapping the playback button.

When the receiver (i.e., experimenter) decided to make a guess, s/he tapped the send button (up arrow). The sound was then played back simultaneously to the imitator and the receiver. If the guess was correct, the green checkmark lit up on both tablets, and a new set of sounds was loaded to both devices. If the guess was not correct, the red x mark lit up, and the players got another try. If the second guess was wrong as well, the next set of sounds was loaded to both devices.

If the receiver needed more information to make a guess, s/he could tap the yellow question mark. The yellow exclamation mark on the imitator's app then lit up to indicate that the receiver needed more input. The server handled the communication between the devices and

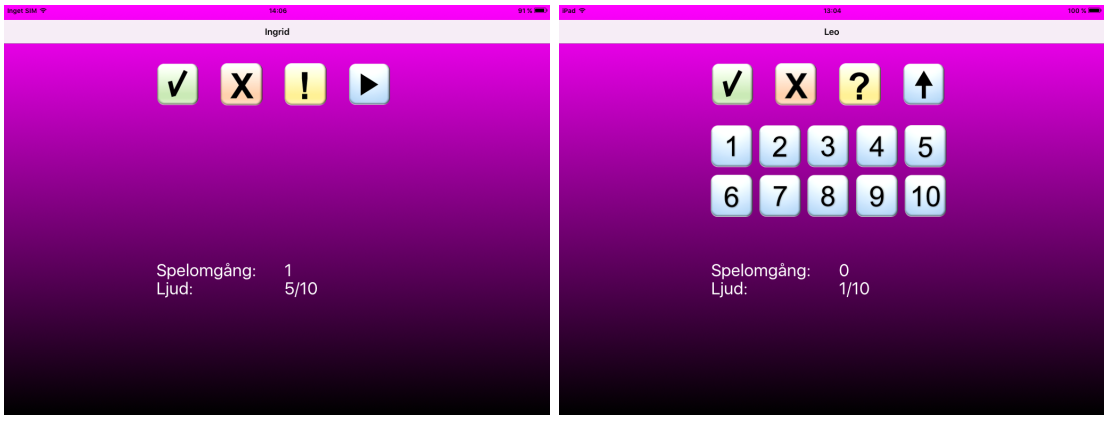

Figure 4: The layout of the imitator's tablet (left) and the receiver's tablet (right).

defined which sounds should be loaded. Each game round consisted of 10 referent sounds to be described. The order of the sounds was randomised for each round of the game, and also for each participant. The server logged and time-stamped all events.

3.4 Imitators

The imitators were recruited via Språkstudion's digital and physical marketing channels. The only requirement for participation was to be a native speaker of Swedish.

Data from 16 imitators were analysed, 6 males and 10 females. The age ranged between 20-70 years old. 14 spoke Central Standard Swedish and 3 had regional colourings, 1 of Gotland, 1 of Jämtland, and 1 of Dalarna. However, one candidate imitator whose native language was Finnish Swedish was rejected since there are prominent differences in the phonological system between Central Standard Swedish and Finnish Swedish. Data recorded for two further imitators was omitted from analysis. One imitator had to be excluded due to apparent problems

with understanding the game instructions. Data for one further imitator was excluded due to technical problems with a recording device.

A bookshop voucher was offered to each imitator that took part in the experiment.

4 Procedure

4.1 Instructions

Experimenter 1 welcomed and introduced the imitator to the experiment. The purpose of the experiment was explained in vague terms to be "research on language learning and sound design". Background data and contact details of the imitator were taken, an informed consent form was signed, and the procedure of the experiment was presented.

The imitator then watched a 3-minute instruction video on how to play the collaborative game. Then, a short training session of the game using two referent sounds was performed. The imitator could ask any questions before starting the recordings.

It should be noted that the first 5 imitators were told that receiver 2 (i.e., experimenter 2, in the second round of the game) did not speak or understand Swedish, English or any other Indo-European language. In the interviews following each recording session, it became increasingly clear that some imitators had difficulties in grasping the notion that a person could "not even understand English". This lead to a change in the introduction to the experiment, and the subsequent imitators were given a more nuanced and concrete cover story, viz. that their next co-player was a native speaker of Georgian and that he was visiting the Department of Linguistics to take part in research. This seemed to be more reassuring for the imitators, and in the post-recording interviews the imitators did not raise the issue the identity of receiver 2.

4.2 Recordings

Continuous audio and video recordings of the imitator were made during the three rounds of the game. Each recording amounted to approximately 30 minutes. The recordings were made using a Røde NT1-A microphone and a Canon XA20 digital video camera.

A technician was monitoring the recording levels and experimental software at all times. An intercom allowed him and the experimenters to communicate with the imitator. The technician also controlled the timing of the appearance of receiver 2 (experimenter 2), who was only present during round 2 of the game in order to minimise the risk of revealing any shared linguistic competence.

After the recordings, the imitators were interviewed briefly about their experience of the experiment. Questions were asked about what they thought was hard or easy, and if they felt they had changed strategies between the different rounds of the game.

5 Annotation

Since imitators differed in how many attempts they made at imitating each referent sound, as well as in how varied the imitations were, selecting which tokens to include in the annotation

was not trivial. First, the referent sounds varied in the number of cycles or events that occurred in the sound (cf. the section on referent sounds above). If all cycles/events in were annotated for each imitation, the dataset would be highly biased towards those referent sounds that had the most cycles/events. Therefore, only one representative cycle/event was annotated for each imitation.

Second, including all attempts for each referent sound in each round would lead to a bias towards those imitators who produce the most tokens. Still, if two or more clearly different imitations were produced for one referent sound, choosing only one token for that referent sound might not lead to an accurate estimation of tameness for that item. To balance these constraints, up to 3 tokens for each referent sound in each round were selected for annotation, but only if the tokens differed significantly with regard to the annotation parameters. Imitations of a total of 437 referent sounds were annotated, of which 366 included 1 imitation token, 67 included two tokens and 4 included three tokens. The total number of annotated imitations was 512. For 21 referent sounds, the imitator did not produce an imitation, but only a verbal description. The recording of a further 22 referent sounds failed due to technical issues.

The data were annotated using the multimodal annotation software ELAN, developed by The Language Archive at the Max Planck Institute for Psycholinguistics, Nijmegen, NL [WBR+06]. The variables annotated were: syllable structure, syllable durations, segmental content and voice. The purpose of the annotation was primarily to obtain an estimate of each utterance along the wild-tame continuum. This estimate is explained further in the following section. An example of the annotation is given in Figure 5.

6 Wild-Tame score – estimating the degree of tameness

To estimate the degree of tameness for imitations a score was constructed based on both prosodic and segmental articulatory aspects. Henceforth we refer to this as W-T score (wild-tame score). Essentially, the W-T score attempts to estimate the degree to which an observed utterance resembles a syllable in Swedish "citation form" pronunciation.

At the outset, all utterances were given the maximal W-T score of 5. The utterances were then penalized for each violation in adherence to linguistic form along three dimensions: syllable structure, voice quality and segmental content.

Syllable structure and duration: An estimate was made as to whether the utterance could be analysed in terms of syllabic elements, for which the minimum requirement was a discernable vocalic nucleus (the presence of a nucleus and a coda were also noted). Phonetically, the vocalic nucleus should be a voiced segment produced with an open oral cavity (i.e., without a consonantal obstruction). Utterances that failed to meet the criteria for syllabic elements were assigned a W-T score of 0 (i.e., a penalty of -5).

For utterances that were judged to be syllabic, the duration of the rime (nucleus + coda) were measured. A penalty was assigned to utterances that deviated from typical rime durations measured for Central Standard Swedish, which tend not to exceed approximately 400 ms [HRS13]. The value of 400 ms was set as the threshold below which no duration penalty would be assigned to the utterance. For durations above 700 ms, a duration penalty of -4 points was assigned to the utterance (the maximum penalty assigned for duration). For du-

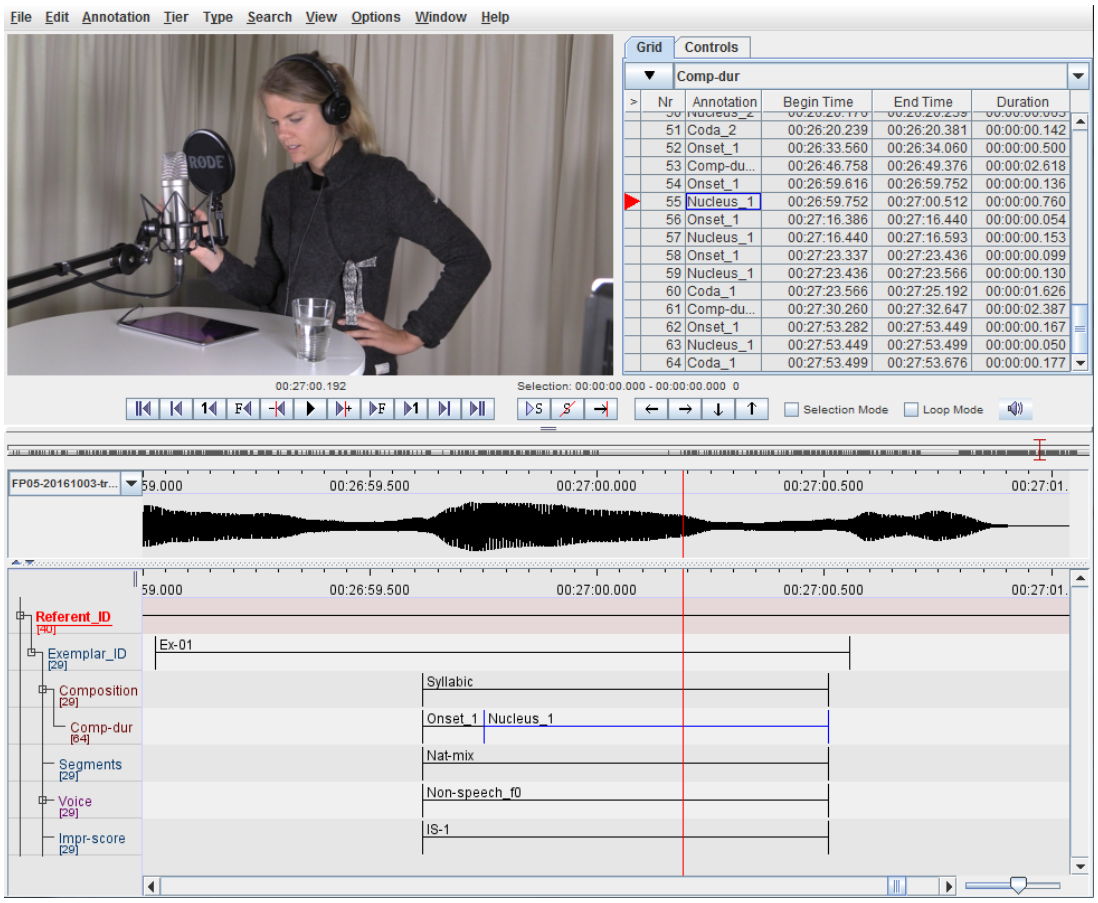

Figure 5: Example of ELAN annotation

Table 2: Penalties assigned for voice deviation

Voice deviation	Penalty
Voiceless	-4
Voice quality, pitch and loudness	-4
Voice quality and pitch	-3.5
Voice quality and loudness	-3.5
Voice quality	-3
Pitch and loudness	-2.5
Pitch	-1.5
Loudness	-1.5

Table 3: Examples of W-T scores for 5 utterances in the data with approximate phonetic transcriptions. An utterance is assigned a W-T score of 5 at the outset, from which various penalties are subtracted. Note that the lowest W-T score for syllabic utterances is 1, even though the accumulated penalties may be more than -4 points, as is exemplified in the second to last item. For asyllabic utterances a W-T score of 0 is always assigned.

Transcription	W-T score	Syllable penalty	Voice penalty	Segment penalty
[drɔp]	5	0 (257 ms rime)	0 (speechlike)	0 (native)
[bɪp]	3.5	0 (130 ms rime)	-1.5 (f0 dev)	0 (native)
[plɪŋŋŋŋŋŋ]	1	-4 (840 ms rime)	0 (speechlike)	0 (native)
[puuummmmm]	1	-4 (1070 ms rime)	-4 (f0+dB+qual)	0 (native)
[$p^w x^w x^w x^w$]	0	-5 (asyllabic)	-4 (voiceless)	-0.5 (mix)

rations between 400 and 700 ms, a penalty was assigned that increased gradually from 0 to -4.

Voice: All utterances were given a voice score based on deviation from the imitators' habitual speaking voice. The voice score was based on judgements of voice quality, loudness and pitch. The principal variations from habitual voice quality were falsetto and epilaryngeal phonation. These constitute a significant deviation from typical use of voice in spoken Swedish, and were assigned a relatively high penalty (see Table 2). Deviations in pitch and loudness were assigned lower penalties. A list of voice deviations and associated penalties is given in Table 2.

Segmental content: The segmental content of the utterances was evaluated in terms of whether they contained fully native segments only, a mixture of native and non-native segments, or non-native segments only. The penalties assigned were 0, 0.5 and 1 respectively.

The total W-T score was calculated by subtracting the duration, voice and segment penalties from the maximal score of 5. A few examples of WT scores with associated penalties, along with approximate phonetic transcriptions are given in Table 3.

The estimates obtained with the W-T score will be used for comparisons across variables within the dataset, for example to compare the tameness of imitations in the three rounds

Table 4: Mean W-T scores, standard deviations and the number of tokens for each of the three test conditions (game rounds).

Test condition	N	Mean	StdDev
1	176	2.14	2.021
2	167	1.12	1.468
3	169	1.08	1.421

of the game, or across different referent sounds. They are not mean to be taken as absolute measures of the degree to which an utterance can be regarded as word-like or onomatopoeic.

7 Results

7.1 Differences across test conditions (game rounds)

The mean W-T score for the three test conditions (i.e., the three rounds of the game) is given in Table 4, along with standard deviation and number of tokens for each condition. The shared language condition (first round of the game) has the highest mean W-T score, 2.14. The mean W-T scores are similar for both non-shared conditions, 1.12 and 1.08 respectively. The distribution of the scores is shown in the box-plot in Figure fig:6. A one-way ANOVA

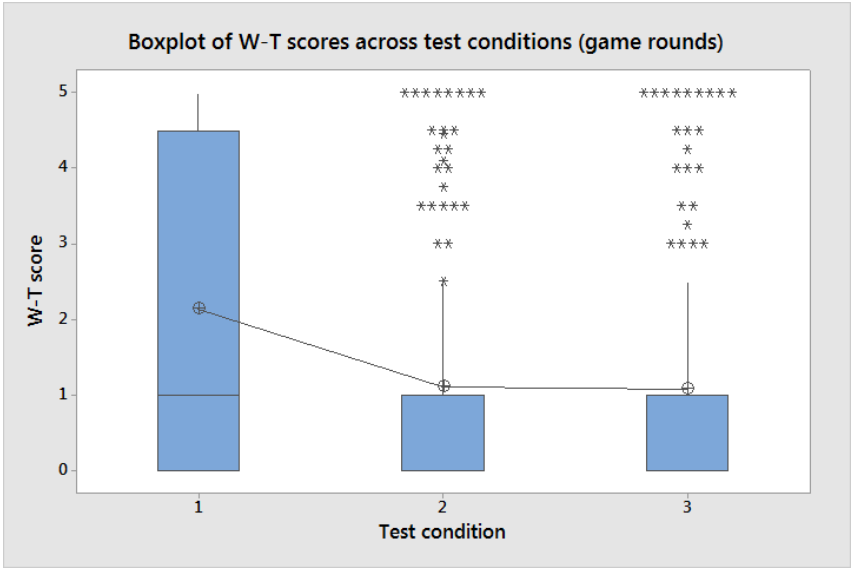

Figure 6: Box-plot showing the distribution of W-T scores for each of the three test conditions (game rounds)

(Analysis of Variance) indicated statistically significant differences for test condition ($F(9,502) = 12.22$, $p < .001$). Pair-wise Tukey's comparisons of the means for the three conditions suggest that the shared language condition is statistically significantly different from the two non-shared conditions, but that the two non-shared conditions are not significantly different

Table 5: Results of a pair-wise Tukey's comparison of means for the three test conditions (game rounds)

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	P-Value
2 - 1	-1.027	0.180	(-1.449; -0.606)	-5.71	0.000
3 - 1	-1.061	0.179	(-1.481; -0.642)	-5.92	0.000
3 - 2	-0.034	0.182	(-0.459; 0.391)	-0.19	0.981

from one another (see Table 5 for p-values). The data for the shared condition exhibit a bimodal distribution of the WT-scores. The histograms in Figure 7 show that most W-T scores in the shared condition are either relatively high (above 3.5) or low (below 1.5), while intermediate values are scarce. For the two non-shared conditions, the histograms are positively skewed, with most of the scores below 1.5 and relatively few intermediate and high scores.

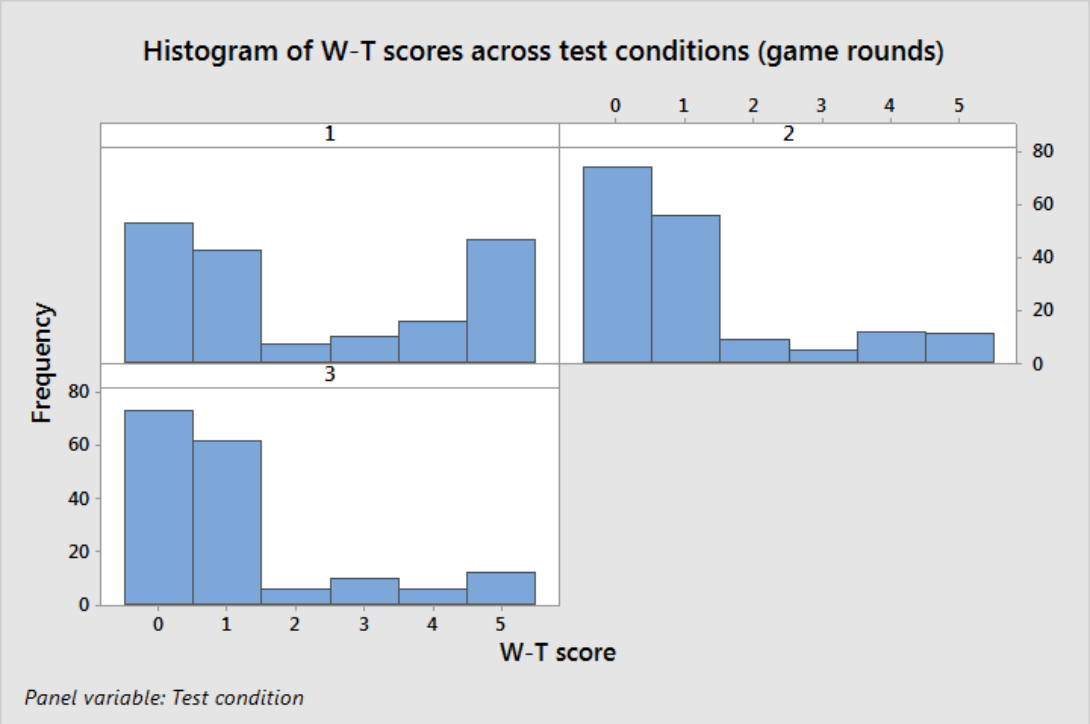

Figure 7: Histograms showing the frequency distribution of W-T scores for each of the three test conditions (game rounds)

7.2 Differences across referent sounds

Mean W-T scores (along with standard deviations and number of tokens) for each referent sound are given in Table 6. There are fairly large differences in mean WT-scores between referent sounds, with mean scores as high as 2.53 for Tick and as low as 0.10 for Flow. A Tukey's pair-wise comparison of means suggests that many of the pair-wise differences are statistically

Table 6: Mean W-T scores, along with standard deviations and number of tokens, for each of the three test conditions (game rounds)

Referent sound	N	Mean	StDev	Grouping
Tick	55	2.53	1.527	A
Alarm	47	2.26	1.191	A B
Drip	48	2.14	1.496	A B C
Click	63	1.83	2.122	A B C D
Bell	62	1.44	2.054	B C D
Saw	45	1.32	0.650	B C D E
Wipers	49	1.18	0.757	C D E
Bom	55	1.07	1.791	D E F
Mixer	45	0.30	2.083	E F
Flow	43	0.10	1.031	F

significant (see Table 6, in which means grouped according to pair-wise significance). Means that do not share a letter are significantly different. Histograms showing the distribution of W-T scores for each referent sound are given in Figure 8. Evidently, some referent sounds yield more bimodal distributions than others, in particular Click, Tick and Drip, while some almost never yield a high W-T score, like Flow and Mixer.

7.3 Inter-imitator differences

Table 7 gives the mean W-T scores for the 16 imitators (for all conditions pooled). The means ranged from 0.61 (FP16) to 2.41 (FP13). A General Linear Model (GLM) analysis of variance of W-T scores with Test condition as a fixed factor, and Referent sound and Imitator as random factors, was performed to examine differences between imitators. It indicated that Imitator is a statistically significant factor ($F(9,485) = 3.00$, $p < .001$), but, as shown in Table 7, a Bonferroni comparison of means did not indicate any clear grouping in W-T scores among the imitators. Figures 9 and 10 show the mean W-T scores for three test conditions (game rounds) for female and male imitators respectively (the split between male and female imitators is done mainly for the sake of exposition, to avoid having too many lines in a single graph). It is apparent that most imitators have a higher W-T score in the shared language condition (test condition 1) than in the non-shared conditions (2 and 3). Only one female (FP12) and one male (FP15) imitator do not follow this pattern. The mean W-T score for female and male imitators is given in Table 8. To test if the difference in mean W-T scores between the sexes was statistically significant, a GLM analysis of variance with Test condition and Sex as fixed factors and Referent sound as a random factor was performed. The GLM analysis indicates no statistically significant difference between males and females ($F(1.499) = 1.18$. $p = .227$).

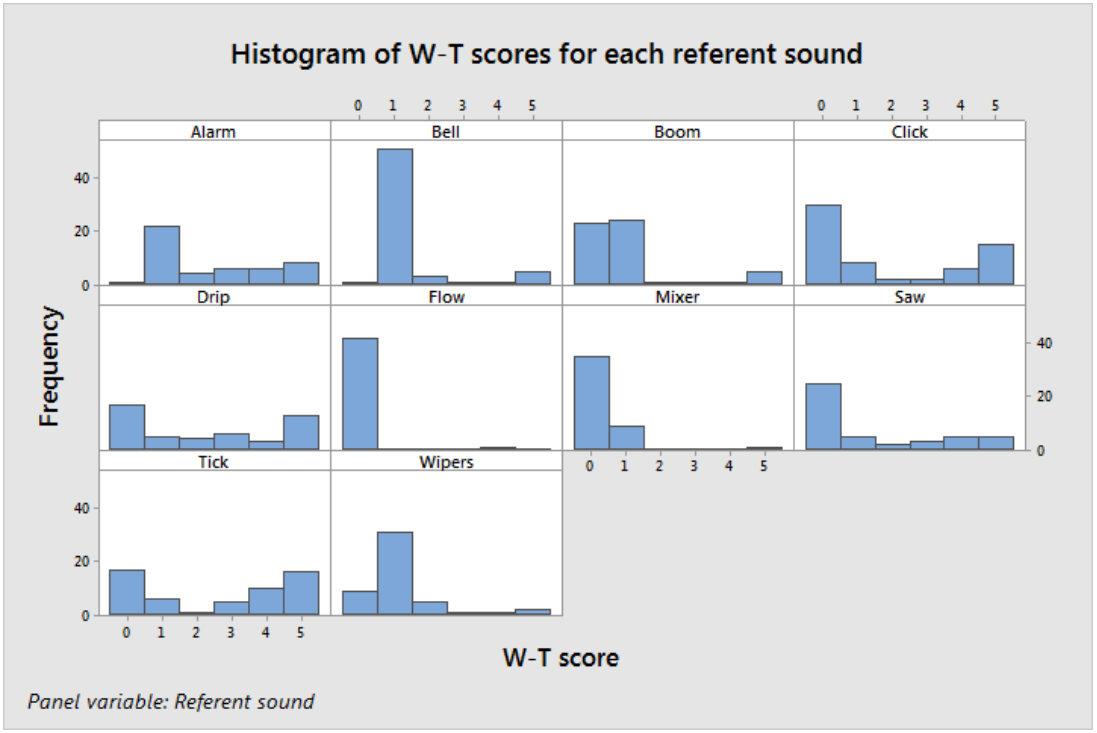

Figure 8: Histograms showing the frequency distribution of W-T scores for each referent sound.

.

Table 7: Mean W-T scores, along with standard deviation and number of tokens, for female and male imitators

Imitator	N	Mean	Grouping
FP13	27	2.41	A
FP05	29	2.11	A B
FP12	29	2.01	A B
FP01	34	1.76	A B C
FP09	31	1.72	A B C
FP04	34	1.52	A B C
FP18	36	1.41	A B C
FP02	33	1.31	A B C
FP10	34	1.34	A B C
FP17	38	1.25	A B C
FP15	33	1.19	A B C
FP03	29	1.18	A B C
FP08	34	1.11	A B C
FP07	27	1.00	A B C
FP14	32	0.91	B C
FP16	32	0.61	C

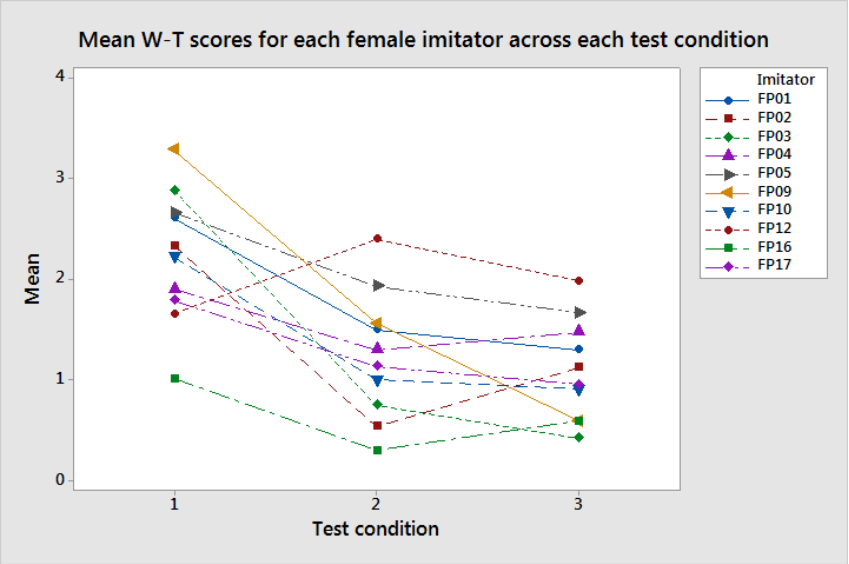

Figure 9: Mean W-T scores across test condition for each female imitator.

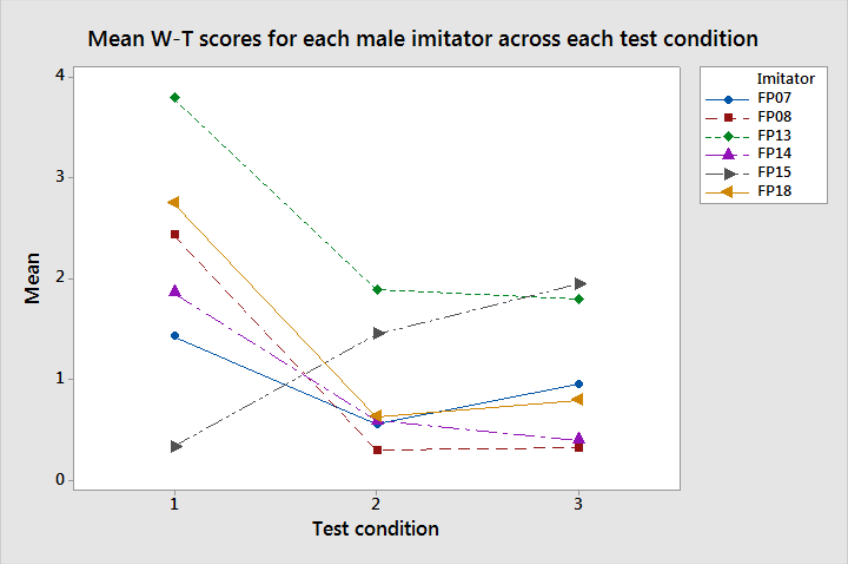

Figure 10: Mean W-T scores across test condition for each male imitator.

Table 8: Mean W-T scores. along with standard deviation and number of tokens. for female and male imitators

Sex	N	Mean	StDev
FEMALE	323	1.504	1.7779
MALE	189	1.378	1.660

8 Discussion and conclusions

It is evident from the present experiment that given the task of describing sounds to a receiver, imitators rarely fail to produce imitative utterances, even if they are not given specific instruction to do so. It stands to reason that this applies to both test conditions with no shared language (game rounds 2 and 3), since having no shared language essentially limits the available expressive means to mimicry and gestures. For the shared language condition, however, the imitators were asked to describe the referent sounds using any expressive means available and were not required to imitate the sound at all. Nevertheless, the imitators rarely use a verbal description only in the shared condition, but usually complement it by some kind of imitative articulation. This indicates that imitation of sound is a natural behaviour that does not require specific instruction or encouragement.

At the higher end of the W-T score range, a substantial difference was observed between imitations produced in the shared linguistic setting of the first game round, as compared to imitations in the two non-shared settings, human as well as computer. When the imitations were embedded in a linguistic context, the imitators produced 47 imitations with a W-T score of 4.5 or higher, compared to 11 and 12 instances in the two non-linguistic contexts respectively. Of the 16 imitators, 14 followed this pattern of producing more tame imitations in the linguistic context.

The most reasonable interpretation of this difference is that the linguistic embedding of imitations greatly increases the likelihood for a tame imitation, i.e., an imitation that is either integrated in their speech flow or highly speechlike. The non-linguistic contexts generate few tame imitations because the imitators are not engaged in a linguistic task while producing the imitations. The nature of the receiver, human or computer, does not seem to be essential, since the results for the two conditions testing for that difference are nearly identical.

The distribution of W-T scores in the shared linguistic setting suggests that the imitators tend to produce either quite wild or tame imitations, rather than imitations that are intermediate between wild and tame. One interpretation of this is that the imitators invoke different articulatory strategies in their utterances, which we can refer to as "language mode" and "sound mode". Language mode involves activating articulatory pathways via a filter of linguistically habitualised articulatory constellations and plans. As a consequence, the utterances are shaped by the articulatory contingencies and hierarchies that apply in speech. In sound mode, by contrast, the imitators utilize the articulatory pathways more directly, apparently by-passing linguistic filters to a large extent. The constraints of speech do not apply and the imitators instead produce imitations that are subject to general articulatory constraints imposed by the human vocal apparatus.

While the shared language condition induced more speechlike imitations than the non-shared conditions, it should be noted that in all conditions the majority of the imitations produced were at the lower end of the W-T score range. A low W-T score implies that imitations have features that are highly divergent from speech. These features may involve longer durations than normal for speech, non-speech voice quality, such as falsetto or growl, and the use of articulations that are not native to (in this case) Swedish. The simplest interpretation of this is that the "sound mode" is the prevalent mode of articulation when it comes to sound imitation. When the imitator becomes engaged in a verbal exchange the chances of invoking language mode, thus producing phonologically constrained imitations,

increases.

However, all referent sounds are not equal when it comes to producing tame imitations. The histogram in Figure 8 (see the results section, above) shows that the referent sounds Flow and Mixer are rarely rendered in a tame fashion (1 instance each). Click and Tick, on the other hand, are frequently rendered through tame imitations.

One reason for this may simply be the fact that click and tick sounds have established onomatopoeic representations in Swedish (as well as in English and a host of other languages). It stands to reason that having an established onomatopoeic form for a particular type of sound creates a bias towards tame imitations for a referent sound of that type. Of course, it would be preferable to try to exclude this factor in the experimental design, but this would eliminate certain basic and commonly occurring types of sounds entirely from the experiment. No transients, for example, could be included, since transients of any kind (be it the clack of an old typewriter, the click of a bottle cap or the tick of a clock) are well represented onomatopoeically. Nor could we use bell-like sounds (impacts with a long decay, whatever the source), since such sounds have established onomatopoeic expressions.

Considering each referent sound used for this experiment, we can note that five of the sounds have fairly well established onomatopoeic expressions: Tick, Click, Drip, Boom and Bell. The remaining five sounds do not: Alarm, Saw, Wipers, Mixer and Flow. Looking to the number of relatively tame renderings of each referent sound, though, it appears that the presence of an onomatopoeic form is not a good predictor of tameness. It is true that the onomatopoeic Click, Tick and Drip are absolute leaders when it comes to tameness. However, the non-onomatopoeic Alarm and Saw are represented by relatively tame imitations more often than Bell and Boom.

We suggest, therefore, that the morphology of the referent sound also needs to be considered. In Figure 11 we have plotted the number of imitations with a W-T score of 3.5 or more against the duration of the cycle or event in the referent sound used (cf. section 3.2, above). We should keep in mind that a key factor in this plot is our estimate of what constitutes speechlike duration in the W-T score. Our duration estimate is based on the rime part of the imitative utterance, and we have set duration criteria that are based on durations measured in spoken Swedish (cf. section 6, above). This entails that at rime durations exceeding 400 ms, imitations get a penalty to their W-T score, and that this penalty increases with duration, maxing out at 700 ms.

Given that we accept this a reasonable estimate of what constitutes speechlike duration in Swedish, the data in Figure 11 suggest that the number of tame imitations decreases as the duration of the imitated cycle or event in the referent sound increases. For example, for Alarm, Saw and Wipers, none of which have specific onomatopoeic representations in Swedish, there is a sharp drop in tameness as referent sound duration increases. Given their relatively long durations, the referent sounds Boom and Bell should have fewer tame instances, but the fact that both have well established onomatopoeic representations may promote the production of more tame instances than duration alone would predict. Sounds with a long, constant intensity profile, like Mixer and Flow, are rarely described by the imitators by using a tame, speechlike utterance.

The data in Figure 11 may even point to a partial explanation for why certain types of sounds are more likely to be represented by established onomatopoeic expressions than others in languages in general. It can be hypothesized that if the duration of the imitated event falls

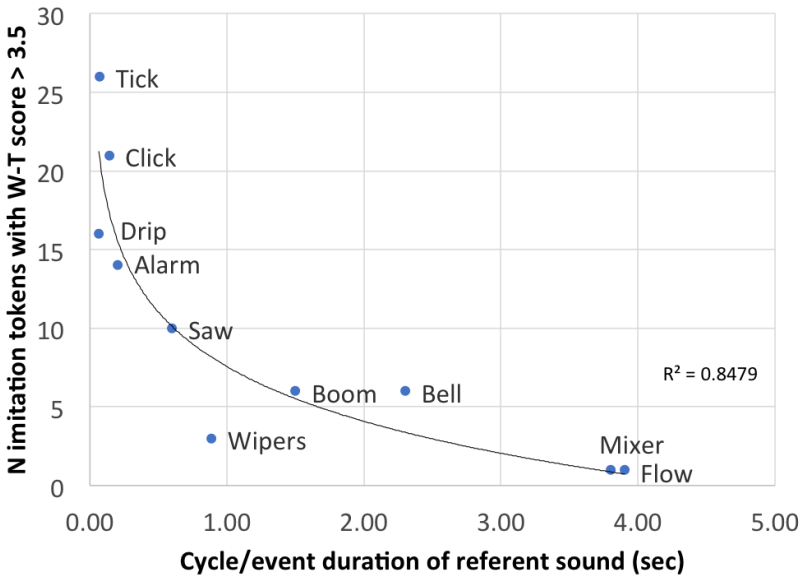

Figure 11: The number of imitations with W-T scores higher than 3.5 plotted against the duration of the imitated cycle/event in each referent sound. A logarithmic regression line with an associated r^2 value of 0.8479 has been fitted to the data.

within the range of typical syllable duration in speech, it is more readily rendered in speechlike form than if its duration extends beyond typical syllable duration. Continuous sounds, like the flow of turbulent gas or liquid should be the least likely to exist in the form of an onomatopoeia. Sounds that fall within the range of typical syllable duration would be more amenable to being rendered in onomatopoeic form. The brevity of some sounds, such as click-like transients, is an inherent and immutable quality of the sound which should serve to make onomatopoeias more likely. Other sounds, such as explosions and ringing impacts, exist in various forms and durations. The Boom and Bell sounds we used for the present experiment were relatively long (1.5 and 2.3 seconds respectively), and are perhaps not typical for the respective class of sounds. The existing onomatopoeic expressions for these types of sounds likely reflects shorter (and more generic) real world referents.

In addition to duration, an essential character of a phonologically well-formed syllable in Swedish (and most other languages) is vocal fold phonation. However, in our data there is no apparent connection between periodicity in the referent sound and tameness in imitation, nor does there seem to be a link between the existence of established onomatopoeic forms and periodicity. Of the referent sounds used for this experiment, four had a distinctive periodic element: Alarm, Wipers, Bell and Mixer. There is little to suggest that the periodicity of these referent sounds has played a role in promoting tameness. Also, despite being aperiodic, transient type sounds, like Tick, Click and Drip, are the ones most frequently rendered using tame imitations in our data and they also have associated onomatopoeic expressions in Swedish (and many other languages). Thus periodicity, as such, does not seem to correlate with the degree of tameness at all.

Summarizing the main results for Task 3.3 and Deliverable 3.3.3, the present experiment implies that imitators use two distinct articulatory strategies for rendering imitations of referent

sounds. One strategy, which we refer to as language mode, is rooted in the imitator's linguistic system and draws on the phonological and phonetic repertoire of the imitator as a speaker. The other strategy, which we refer to as sound mode, reflects the general articulatory capacity of the imitator, unconstrained by the articulatory constraints and filters imposed by the native phonological system and phonetic expression. Sound mode is, instead, constrained by the limitations of the human vocal apparatus as a sound generator. In this context, we should keep in mind that even if imitations often contain articulations that are a good match for linguistically native ones, this should not be taken as evidence for the application of native linguistic structures in imitations, but rather as the result of general articulatory constraints which yield a set of possible articulations, some of which are shared by the two modes of production, language mode and sound mode.

Of the two modes, sound mode seems to be the dominant one and as a results the data set as a whole contains far more wild imitations than tame ones. The tame imitations engendered by language mode occur predominantly when the imitator is engaged in giving verbal descriptions of referent sounds. With the linguistic context removed, the imitators adopt a sound mode of articulation, and are seemingly free to explore their phonetic (or anthropophonic) space and adapt their prosodic patterns to that of the sound being imitated.

For the SkAT-VG system, the implication must be the fundamental assumption that the default mode of the user is sound mode. The SkAT-VG user is not engaged in a linguistic task with an interlocutor capable of interpreting and understanding verbal descriptions of sounds. Therefore, one should not expect SkAT-VG users to spontaneously make use of onomatopoeic (or tame) imitations, and there is no indication that constructing a system of linguistic aliases as alternatives for the input of certain types of sounds would be regarded as helpful or appropriate by the user.

Looking beyond SkAT-VG, the finding that the range from wild to tame imitations seems to be discrete rather than continuous is intriguing and may have implications for research into language education, specifically with regard to pronunciation learning and teaching. There is a long held conception in linguistics, expressed in an influential text by Roman Jakobson [Jak41], that humans are born with a universal articulatory toolbox, evident during the first stages of babbling. The articulatory repertoire then grows more constrained and language-specific as the language is acquired. A large body of research has furnished us with evidence that infants tune in to perceptual categories relevant to the ambient language as early as at six months of age (cf., e.g., [KWL+92]), and we know that by the age of 5 children have, by and large, acquired the specific articulatory constellations and targets of their native language. At a certain point, however, the capacity to acquire a new language and achieve native speaker (L1) competence expires, and, for most people, true L1 competence becomes virtually impossible to achieve.

When it comes to pronunciation, the effects of the native language on a second language (L2), in the form of a foreign accent, are all too obvious. L2 learners tend to apply the articulatory repertoire of their native language to what they view as reciprocal articulations in the L2. Seen in the light of our experimental results, it seems as if L2 learners are often stuck in a language mode of articulation. We are now asking whether the pronunciation of L2 learners might benefit from their being induced to approach the L2 articulation from a sound mode of articulation. This might take the form of a series of exercises that use pure sound imitation as a point of departure, working gradually towards linguistic forms. At present, we are designing an experiment in which we attempt to assess the differences between imitating

sound on the one hand, and producing various types linguistic forms on the other. Hopefully, this will give us a clearer picture of the difference between the two modes and the potential application of the articulation mode concept in L2 acquisition.

References

[And98] Earl R Anderson. *A grammar of iconism*. Fairleigh Dickinson Univ Press, Madison, NJ, 1998.

[Aus94] Robert Austerlitz. Finnish and Gilyak sound symbolism – an interplay between system and history. In L. Hinton, J. Nichols, and J.J. Ohala, editors, *Sound symbolism*, pages 249–260. 1994.

[BWH+16] Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823, 2016.

[Dif94] G. Diffloth. i:big, a:small. In L. Hinton, J. Nichols, and J.J. Ohala, editors, *Sound symbolism*, pages 107–114. 1994.

[Dok54] Clement Martyn Doke. *The southern Bantu languages*. Oxford University Press, London, UK, 1954.

[Ham98] Shoko Hamano. *The Sound-Symbolic System of Japanese.* Center for the Study of Language and Information, Stanford University, Stanford, CA, 1998.

[HRS13] Pétur Helgason, Catherine Ringen, and Kari Suomi. Swedish quantity: Central standard swedish and fenno-swedish. *Journal of Phonetics*, 41(6):534 – 545, 2013.

[Jak41] Roman Jakobson. *Kindersprache, Aphasie und allgemeine Lautgesetze*. Almqvist & Wiksell, Uppsala, Sweden, 1941.

[KWL+92] Patricia K Kuhl, Karen A Williams, Francisco Lacerda, Kenneth N Stevens, and Bjorn Lindblom. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255:606–608, 1992.

[WBR+06] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of LREC – Fifth International Conference on Language Resources and Evaluation*, pages 1556–1559, 2006.