A database of articulatory annotations of vocal imitations

Pétur Helgason[1,2], Gláucia Laís Salomão[1,2], Sten Ternström[2]
[1] Department of Linguistics, Stockholm University, Sweden
[2] Dept. of Speech, Music and Hearing, Royal Institute of Technology, Sweden
petur.helgason@ling.su.se, gsalomao@kth.se

Abstract

This paper gives a short overview of a part of a database of vocal imitations of sounds collected within the project SkAT-VG: Sketching Audio Technologies using Vocalizations and Gestures, which have been annotated for articulation. The data comprise video and audio recordings of vocal imitations of 50 referent sounds produced by four Swedish improvisational actors. Eight articulatory parameters were annotated. ELAN, a software tool for annotating and managing video and audio data, is used as a container for the data base.

Introduction

In sound design activities, communicating and describing sonic ideas can be problematic. In this context it is common to use vocal and gestural imitations to convey the mental image of a sound, and vocal imitations have been shown to be more effective than verbal communication in conveying acoustic information to listeners (Lemaitre and Rochesso, 2014).

The SkAT-VG project aims to construct a design tool that enables sound designers to use their own vocal apparatus for sound sketching, i.e., in the very early stage of the sound design process. For this purpose, an investigation of which articulatory mechanisms imitators can make of when imitating sounds was needed. A part of the data gathered within the project for this purpose is presented here.

Data acquisition

Four imitators were recorded, all Swedish speaking improvisational actors, recruited through an agency and paid for their participation. The imitators were aged 20-40, two male and two female. The recordings were carried out in a soundproofed booth. An audio signal, an electroglottographic (EGG) signal and a video feed from two angles (cf. Figure 1) were recorded. The referent sounds (i.e. the sounds that were to be imitated) were presented to the imitators through a loudspeaker. The order of the referent sounds was randomised and they were presented and recorded in groups of 10. Each such group typically yielded a 15 to 20 minute recording. For each group of referent sounds, the imitators were able to listen to the referent sounds, as well as their own imitations at will and they could revisit referent sounds and redo their imitation of that sound if they wanted. The last imitation they performed for a given referent sound was the one that was subjected to articulatory analysis.

The referent sounds were selected from 3 major classes (cf. Lemaitre & Heller, 2013): (1) basic mechanical interactions (with solids, liquids and gases as subclasses); (2) abstract sounds; and (3) engine sounds. We strived to achieve a balance between different articulatory mechanisms used by the imitators, but given the unpredictable nature of the imitations, such a balance could not be guaranteed.

The database

Database source files

For the annotation and extraction of data, the database uses ELAN (EUDICO Linguistic Annotator), a freeware database tool produced by The

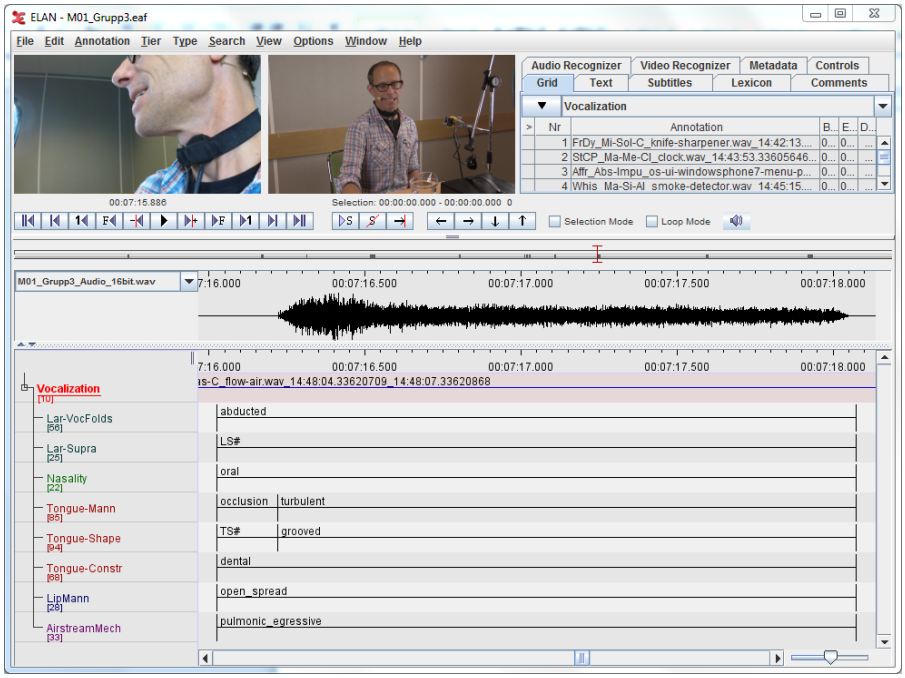

Figure 1: ELAN screen layout with annotation example 1. Side and front camera views are at top left. At the top right is the list of indexing labels. At the bottom are the eight articulatory annotation tiers. The referent sound is “gas squeezing through a narrow aperture” and the imitation sounds roughly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

Language Archive project at the Max Planck Institute for Psycholinguistics in Nijmegen. ELAN is an annotation tool for creating, editing and extracting data from multi-media recordings.

The media sources for each ELAN file in the database comprise 2 video recordings, an audio recording and an electroglottographic (EGG) recording (the latter is not available for all subjects, though). In order to preserve the context in which imitations were performed (preceding and following imitation attempts), entire sessions are included in the ELAN files (rather than individual imitations). There are normally 5 ELAN data files per imitator, each comprising 10 annotated imitations. The data files vary in duration, but are typically between 15 and 20 minutes long.

Annotation

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 consists of 8 annotation tiers, each representing a specific articulatory parameter. Two of the tiers describe articulatory actions in the larynx, three tiers describe the actions of the tongue (both tongue body and tongue tip), one tier is devoted to the lips, one tier controls for nasality and, finally, one tier describes the airstream mechanism (or sound initiation).

The resulting annotation in effect comprises an articulatory score that describes the contribution or action of individual articulators over time, and gives a holistic description of the articulatory mechanisms used for the imitation. The annotation is rich in articulatory detail to ensure that all aspects that may be significant for conveying referent sounds through imitation are covered.

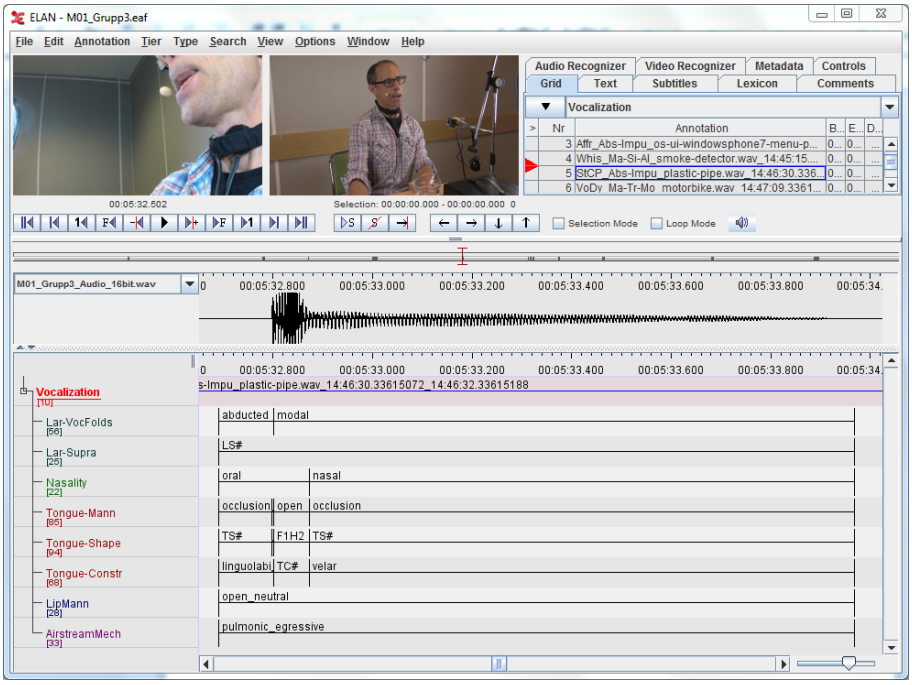

Figure 2: ELAN screen layout with annotation example 2. The referent sound is “plastic pipe being struck against a solid object” and the imitation sounds roughly like “ding”.

The following section gives some examples that serve to illustrate the annotation parameters.

Example 1

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, 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 eight articulation tiers were combined into one, only two distinct articulatory segments would emerge. The vocal folds are **abducted** throughout (Lar-VocFolds tier) 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 (a [t̪] in IPA terms), but then the closure is released and a dental **turbulence** is created with a **grooved** tongue (Tongue-Shape). This essentially results in a [s̪]-sound, which is maintained for almost two seconds.

The articulatory annotation in this case follows IPA sound descriptors quite closely: [t̪] is a pulmonic egressive, voiceless, dental, oral stop. The [s̪] is, similarly, a pulmonic egressive, voiceless, (lamino-)dental, grooved, oral fricative.

Example 2

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. Here, one can distinguish a total of four 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 (less than 10 ms), turbulence is created as the tongue tip parts with the upper lip (the relevant annotations are not readable at the zoom level of the screenshot in Figure 2). At the end of the release, the vocal folds produce phonation (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].

Data retrieval

ELAN allows data extraction in several formats, including simple tab-delimited text files and Praat TextGrids. ELAN also supports merging tiers to create combinatory annotation tiers.

The annotation system was designed with scalability in mind, i.e. that it should be easy to examine and retrieve the actions of different articulators separately, or in any combination of choice. For example, the actions of the lips (LipMann) can be retrieved independently from other articulations, or in combination with other parameters (e.g. Tongue-Mann and Tongue-Shape). Similarly, information about tongue shape and place of constriction can be disregarded while information on tongue manner is retained.

Discussion

The database has yielded a wealth of information about the articulatory mechanisms employed in doing vocal imitations of sounds. A crucial finding is that the improvisational actors we recorded do not seem constrained by the articulatory mechanisms of their native language. Instead, they were observed to use various "exotic" types of sound initiation (e.g. glottalic egressive and velaric ingressive), as well as combinations of manner and place of articulation that do not occur in Swedish (e.g. bilabial trills and linguolabial stops).

The database is intended to be accessible to all interested researchers. Given the size of the source files (particularly the video files), making a version of the database available online is difficult but we are currently looking into ways of making this possible.

Acknowledgements

This research was conducted within the SkAT-VG project (2014–2016), financed by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 1096 618067.

References

Lemaitre, G. and Rocchesso, D. (2014). "On the effectiveness of vocal imitation and verbal descriptions of sounds," *Journal of the Acoustical Society of America 135*, 862–873.

Lemaitre, G., and Heller, L. M. (2013). "Evidence for a basic level in a taxonomy of everyday action sounds," *Exp. Brain Res. 226*, 253–264.