# Classifying non-speech vocalisations for speaker recognition

*Finnian Kelly*[1]*, Harry Swanson*[2]*, Kirsty McDougall*[2]*, and Anil Alexander*[1]
[1]*Oxford Wave Research Ltd., Oxford, U.K.*
{finnian|anil}@oxfordwaveresearch.com
[2]*University of Cambridge, Cambridge, UK.*
{hs686|kem37}@cam.ac.uk

Non-speech vocalisations (NSVs) are sounds speakers can produce with their vocal organs that do not have linguistic content, and that may or may not contribute meaning to a communication. Among such sounds are laughter, screams, yawns, moans, groans, sighs, throat clearings, hiccups, sneezes, and paralinguistic clicks.

There is little existing research on the relevance of NSVs to forensic speaker recognition, and in automatic speaker recognition they are typically discarded by the voice activity detection process, which occurs prior to speaker modeling and comparison. However, there have been some promising findings, e.g. Bachorowski et al. (2001) used laughter to classify speakers at above-chance levels using an automatic approach, and Engelberg et al. (2019) found participants were able to discriminate between speakers at above-chance levels from scream stimuli.

Despite the very limited research base, forensic practitioners do examine and sometimes use NSVs in real casework. Gold and French's survey of 36 FSC practitioners noted that 94% of respondents reported "examining non-linguistic features at least some of the time" (2011:302). This study explores whether real examples of NSVs can be classified automatically, with the aim of assessing their speaker-characterising properties, and ultimately of informing their use in speaker recognition.

Anikin & Persson's (2017) corpus of spontaneous NSVs (N = 603) was used as a source of data. The corpus comprises audio clips extracted from YouTube videos, containing either a single syllable or a bout (series of syllables) produced in a single emotional state. Each clip is labelled with one of nine emotional categories, and one of eight call types (grunt, laugh, moan, roar, scream, sigh, tone, whimper). Anikin & Persson's corpus is favourable to other NSV corpora (Belin et al. 2008, Sauter et al. 2010, Lima et al. 2013, Holz et al. 2021) as it contains spontaneous, rather than acted vocalisations.

A pilot classification study was conducted using VOCALISE x-vectors (Kelly et al., 2019) within a speaker profiling framework. Audio clips from four classes of NSV call types were selected: roar (N=84), scream (N=91), laugh (N=109), and moan (N=38). Additionally, a 'normal' speech class (N=100) was created by extracting short audio clips of spontaneous speech from YouTube videos. All NSV and speech clips came from different speakers. A two-class experiment was conducted, whereby a classifier was trained and tested for every possible pairwise combination of the five classes (4 NSV, 1 speech). In each case, recordings were randomly divided into training and testing sets (ratio 3:1). A linear support vector machine (SVM) was trained using spectral (MFCC) x-vectors extracted from the training set, and applied to classify x-vectors extracted from the testing set. This was repeated 10 times with a different random train-test partition. The resulting average classification EERs (Equal Error Rates) were <1% for all combinations of NSV vs speech, and between 5.6% (laugh vs roar) and 11% (scream vs roar) within NSVs. This promising discrimination performance supports the use of spectral x-vectors for NSV classification, which will enable a systematic assessment of the effects of NSVs on speaker recognition.

## References

Anikin, A., & Persson, T. (2017). Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. Behavior research methods, 49(2), 758-771.

Bachorowski, J. A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. The Journal of the Acoustical Society of America, 110(3), 1581-1597.

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods, 40, 531–539. doi:10.3758/BRM.40.2.531

Engelberg, J. W., Schwartz, J. W., & Gouzoules, H. (2019). Do human screams permit individual recognition? PeerJ, 7, e7087.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. International Journal of Speech, Language and the Law, 18(2), 293-307.

Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2021). The paradoxical role of emotional intensity in the perception of vocal affect. Scientific reports, 11(1), 1-10.

Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep Neural Network Based Forensic Automatic Speaker Recognition in VOCALISE using x-Vectors, 2019 AES International Conference on Audio Forensics.

Lima, C. F., Castro, S. L., & Scott, S. K. (2013). When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing. Behavior Research Methods, 45, 1234–1245.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. Quarterly Journal of Experimental Psychology, 63(11), 2251-2272.