This work deals with the task of melody extraction from symphonic music recordings, where the term “melody” is understood as “the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music and that a listener would recognise as being the ‘essence’ of that music when heard in comparison”. Melody extraction algorithms are commonly evaluated by comparing the pitch sequences they estimate against a "ground truth" created by humans. In order to collect evaluation material from our target repertoire, classical music excerpts in large ensemble settings, we collected recordings of people singing along with the music. In this work, we analyse such recordings and the output of state-of-the-art automatic melody extraction methods, in order to study the agreement between humans and algorithms. Agreement is assessed by means of standard measures that compare pitch sequences on a frame basis, mainly chroma accuracy, which ignores octave information. We also study the correlation between this agreement and the properties of the considered musical excerpts (e.g. melodic density, tessitura, complexity) and of the subjects (e.g. musical background, degree of knowledge of each piece). We confirm the challenges of melody extraction for this particular repertoire, and we identify note density and pitch complexity as the melodic features most correlated to the accuracy and mutual agreement for both humans and algorithms.