# Natural Language Processing for Mixed Speech-Music Playlist Generation

Ivett Benyeda[1], Mátyás Jani[2], Gergely Lukács[2]

[1] Research Institute for Linguistics, Hungarian Academy of Sciences
33, Benczúr str., Budapest, HU-1068, Hungary
benyeda.ivett@nytud.mta.hu
[2] PPCU Faculty of Information Technology and Bionics
50/A, Práter str., Budapest, HU-1083, Hungary
{jani.matyas, lukacs}@itk.ppke.hu

## Abstract

Music listening habits are changing with the spread of online media consumption and the usage of smartphones. Large online music collections have become available and there is a need for selecting and ordering pieces of music automatically, for a customised listening experience. This process, the playlist generation, has gained much research attention recently and got implemented recently in popular music streaming services. The mainstream focuses on the acoustics of the playlist generation. Some current studies have revealed that natural language processing can also improve the results, especially in the mood detection of the songs. These approaches focus on music only playlists.

Mixed speech-music playlists are different from those in the approach that they contain audio recordings with speech (interviews, actual news, etc.) alongside with musical pieces. Such playlists allow new, innovative applications, through which users can listen to music matching their tastes, and they are also connected with the external world and actual events. The first approaches on mixed speech-music playlists focused on the acoustics of the audio clips.

In this paper preliminary experiments are presented towards the generation of mixed speech-music playlists with the help of language technology, an earlier untouched area. In our work, first the relevant connecting points between recordings containing speech and music pieces were examined with the help of professional radio editors. This revealed that the most important connecting points are (1) the mood of the parts and in some cases, especially in the case of feasts (2) the matching of topics.

The most straightforward natural language processing approaches for both parts are to use special mood and feast lexicons. Experiments were conducted based on English language radio podcasts and on their transcripts. A major challenge is that automatic speech recognition (ASR) technologies are required to produce the transcripts. ASR can be used either to recognise the whole speech, this is the so called spoken term detection, or only to recognise some selected keywords, the so called keyword search. Our experiments on a limited dataset using an ASR system suggest that the limited quality achievable with ASR does not affect significantly the quality of the mood detection.