Abstract

This thesis proposes specific signal processing and machine learning methodologies for automatically aligning the lyrics of a song to its corresponding audio recording. The research carried out falls in the broader field of music information retrieval (MIR) and in this respect, we aim at improving some existing state-of-the-art methodologies, by introducing domain-specific knowledge.
The goal of this work is to devise models capable of tracking in the music audio signal the sequential aspect of one particular element of lyrics - the phonemes. Music can be understood as comprising different facets, one of which is lyrics. The models we build take into account the *complementary context* that exists around lyrics, which is any musical facet complementary to lyrics. The facets used in this thesis include the structure of the music composition, structure of a melodic phrase, the structure of a metrical cycle. From this perspective, we analyse not only the low-level acoustic characteristics, representing the timbre of the phonemes, but also higher-level characteristics, in which the complementary context manifests. We propose specific probabilistic models to represent how the transitions between consecutive sung phonemes are conditioned by different facets of complementary context.
The complementary context, which we address, unfolds in time according to principles that are particular of a music tradition. To capture these, we created corpora and datasets for two music traditions, which have a rich set of such principles: Ottoman Turkish makam and Beijing opera. The datasets and the corpora comprise different data types: audio recordings, music scores, and metadata. From this perspective, the proposed models can take advantage both of the data and the music-domain knowledge of particular musical styles to improve existing baseline approaches.
As a baseline, we choose a phonetic recognizer based on hidden Markov models (HMM): a widely-used methodology for tracking phonemes both in singing and speech processing problems. We present refinements in the typical steps of existing phonetic recognizer approaches, tailored towards the characteristics of the studied music traditions. On top of the refined baseline, we device probabilistic models, based on dynamic Bayesian networks (DBN) that represent the relation of phoneme transitions to its complementary context. Two separate models are built for two granularities of complementary context: the structure of a melodic phrase (higher-level) and the structure of the metrical cycle (finer-level). In one model we exploit the fact the syllable durations depend on their position within a melodic phrase. Information about the melodic phrases is obtained from the score, as well as from music-specific knowledge.Then in another model, we analyse how vocal note onsets, estimated from audio recordings, influence the transitions between consecutive vowels and consonants. We also propose how to detect the time positions of vocal note onsets in melodic phrases by tracking simultaneously the positions in a metrical cycle (i.e. metrical accents).
In order to evaluate the potential of the proposed models, we use the lyrics-to-audio alignment as a concrete task. Each model improves the alignment accuracy, compared to the baseline, which is based solely on the acoustics of the phonetic timbre. This validates

our hypothesis that knowledge of complementary context is an important stepping stone for computationally tracking lyrics, especially in the challenging case of singing with instrumental accompaniment.
The outcomes of this study are not only theoretic methodologies and data, but also specific software tools that have been integrated into Dunya - a suite of tools, built in the context of CompMusic, a project for advancing the computational analysis of the world's music. With this application, we have also shown that the developed methodologies are useful not only for tracking lyrics, but also for other use cases, such as enriched music listening and appreciation, or for educational purposes.

keywords:
signal processing
machine learning
music information retrieval
singing voice
lyrics
lyrics-to-audio alignment
phonemes
music scores
Turkish makam music
Beijing Opera
hidden Markov models
Dynamic Bayesian Networks

Resum

La tesi aquí presentada proposa metodologies d'aprenentatge automàtic i processament de senyal per alinear automàticament el text d'una cançó amb el seu corresponent enregistrament d'àudio. La recerca duta a terme s'engloba en l'ampli camp de l'extracció d'informació musical (Music Information Retrieval o MIR). Dins aquest context la tesi pretén millorar algunes de les metodologies d'última generació del camp introduint coneixement específic de l'àmbit.
L'objectiu d'aquest treball és dissenyar models que siguin capaços de detectar en la senyal d'àudio l'aspecte seqüencial d'un element particular dels textos musicals; els fonemes. Podem entendre la música com la composició de diversos elements entre els quals podem trobar el text. Els models que construïm tenen en compte el *context complementari* del text. El context són tots aquells aspectes musicals que complementen el text, dels quals hem utilitzat en aquest tesi: la estructura de la composició musical, la estructura de les frases melòdiques i els accents rítmics. Des d'aquesta prespectiva analitzem no només les característiques acústiques de baix nivell, que representen el timbre musical dels fonemes,

sinó també les característiques d'alt nivell en les quals es fa patent el context complementari. En aquest treball proposem models probabilístics específics que representen com les transicions entre fonemes consecutius de veu cantanda es veuen afectats per diversos aspectes del context complementari.

El context complementari que tractem aquí es desenvolupa en el temps en funció de les característiques particulars de cada tradició musical. Per tal de modelar aquestes característiques hem creat corpus i conjunts de dades de dues tradicions musicals que presenten una gran riquesa en aquest aspectes; la música de l'opera de Beijing i la música makam turc-otomana. Les dades són de diversos tipus; enregistraments d'àudio, partitures musicals i metadades. Des d'aquesta prespectiva els models proposats poden aprofitar-se tant de les dades en si mateixes com del coneixement específic de la tradició musical per a millorar els resultats de referència actuals.
Com a resultat de referència prenem un reconeixedor de fonemes basat en models ocults de Markov (Hidden Markov Models o HMM), una metodologia abastament emprada per a detectar fonemes tant en la veu cantada com en la parlada. Presentem millores en els processos comuns dels reconeixedors de fonemes actuals, ajustant-los a les característiques de les tradicions musicals estudiades. A més de millorar els resultats de referència també dissenyem models probabilistics basats en xarxes dinàmiques de Bayes (Dynamic Bayesian Networks o DBN) que respresenten la relació entre la transició dels fonemes i el context complementari. Hem creat dos models diferents per dos aspectes del context complementari; la estructura de la frase melòdica (alt nivell) i la estructura mètrica (nivell subtil). En un dels models explotem el fet que la duració de les síl·labes depén de la seva posició en la frase melòdica. Obtenim aquesta informació sobre les frases musical de la partitura i del coneixement específic de la tradició musical. En l'altre model analitzem com els atacs de les notes vocals, estimats directament dels enregistraments d'àudio, influencien les transicions entre vocals i consonants consecutives. A més també proposem com detectar les posicions temporals dels atacs de les notes en les frases melòdiques a base de localitzar simultàniament els accents en un cicle mètric musical.

Per tal d'evaluar el potencial dels mètodes proposats utlitzem la tasca específica d'alineament de text amb àudio. Cada model proposat millora la precisió de l'alineament en comparació als resultats de referència, que es basen exclusivament en les característiques acústiques tímbriques dels fonemes. D'aquesta manera validem la nostra hipòtesi de que el coneixement del context complementari ajuda a la detecció automàtica de text musical, especialment en el cas de veu cantada amb acompanyament instrumental.

Els resultats d'aquest treball no consisteixen només en metodologies teòriques i dades, sinó també en eines programàtiques específiques que han sigut integrades a Dunya, un paquet d'eines creat en el context del projecte de recerca CompMusic, l'objectiu del qual és promoure l'anàlisi computacional de les músiques del món. Gràcies a aquestes eines demostrem també que les metodologies desenvolupades es poden fer servir per a altres aplicacions en el context de la educació musical o la escolta musical enriquida.

keywords:
processament de senyal
l'extracció d'informació musical
aprenentatge automàtic
veu de cant
el text del la cançó
alineament de text amb àudio
fonemes
partitura
música makam turc-otomana
opera de Beijing
models ocults de Markov
xarxes dinàmiques de Bayes