ACOUSTICS2008/2211
Modeling Vocal Sounds in Polyphonic Musical Audio Signals

Masataka Goto and Hiromasa Fujihara
National Institute of Advanced Industrial Science and Technology (AIST), IT, 1-1-1 Umezono, Tsukuba, 305-8568 Ibaraki, Japan

This paper describes our research aimed at modeling vocal sounds (singing voices) in available music recordings and its applications to singer identification, singer similarity, and lyrics synchronization. Our predominant-F0 estimation method, *PreFEst*, can obtain the melody line by modeling the input sound mixture as a weighted mixture of *harmonic-structure tone models* (probability density functions) of all possible F0s and estimating their weights and the tone models by MAP estimation. Since the PreFEst was designed for general melodies, we extended it to specialize in vocal melodies by using vocal timbre models — *vocal and non-vocal GMMs.* Those GMMs are trained beforehand and used to evaluate the vocal probability. The GMMs are also used to identify vocal regions, but its strategy should depend on applications. For singer identification and singer-similarity calculation, since the purpose is to model singer's identity by training *each singer's vocal GMM*, certainly reliable vocal regions should be identified even if most true regions were missed. On the other hand, for lyrics synchronization, since the purpose is to align each phoneme to the estimated vocal melody, vocal regions should be identified without missing any true regions. We achieved this by biasing log likelihoods provided by vocal and non-vocal GMMs.