

Mark Hasegawa-Johnson  
Assistant Professor  
ECE Department  
University of Illinois  
Urbana, IL 61801

November 5, 2002

Professor Charles F. Zukoski  
Vice Chancellor for Research  
University of Illinois at Urbana-Champaign  
420 Swanlund Administration Building, MC-304  
601 East John Street  
Champaign, IL 61820

Dear Professor Zukoski,

I would like to express my support for Professor Stephen Downie's proposal for the creation of a Music Information Retrieval development and testing database.

I am currently principal investigator of an NSF-funded grant for "Landmark-Based Speech Recognition in Music and Speech Backgrounds." Speech in films and on TV tends to be recorded over a background of music. Standard speech recognizers ignore the complex temporal structure of the background music, and consequently fail to correctly transcribe broadcast news and film audio. Psychophysical evidence suggests that humans are able to recognize speech in music because we are able to pre-consciously recognize the rhythmic and harmonic structures of both signals, and that we use this information to pre-consciously separate the music and speech into independent audio "streams" for conscious processing. Mimicking human competence in an automatic system requires relatively detailed statistical models of the rhythm and harmonic structures of both speech and music.

Every acoustic recognition model is a model of human perceptual competence. A human making a typical perceptual decision is able to call on the trained parameter settings of billions of synapses, trained through thousands of hours of integrated sensory perception. The number of trainable parameters in an artificial intelligence model, by contrast, is limited by the amount of publicly available training data. Every time the amount of training data doubles, researchers find meaningful ways to double the number of trained parameters in our statistical models; every time we double the number of adequately trained parameters, the performance of the speech recognizer improves.

The Linguistic Data Consortium at the University of Pennsylvania publishes standardized speech training databases containing a few hundred hours of training data; a typical speech recognition model trained using one of the larger databases may contain one million parameters. The field of Music Information Retrieval has been handicapped, in

comparison, by limitations on the amount of standardized publicly available training data. My own attempts to build statistical models of music, for the purpose of recognizing speech in music backgrounds, have been handicapped by limitations on the amount of standardized publicly available training data.

Professor Downie's proposal seeks to make a large standardized database available to the research community in Music Information Retrieval. I believe that the proposed database will also be of use to psychologists, computer scientists, and engineers building statistical models of language and auditory perception. The proposed data will provide a unique and much-needed resource for research in these areas.

Collegially yours,

Mark Hasegawa-Johnson  
ECE Department  
University of Illinois