The Orchestra of Speech: a speech-based instrument system

Daniel Formo
Department of Music
Norwegian University of Science and Technology
Trondheim, Norway
daniel.formo@ntnu.no

ABSTRACT
The Orchestra of Speech is a performance concept resulting from a recent artistic research project exploring the relationship between music and speech, in particular *improvised music* and *everyday conversation*. As a tool in this exploration, a digital musical instrument system has been developed for "orchestrating" musical features of speech into music, in real time. Through artistic practice, this system has evolved into a personal electroacoustic performance concept.

Author Keywords
Speech and music, prosody, improvisation, machine learning, instrument design, artistic research.

CCS Concepts
• **Applied computing → Sound and music computing**; Performing arts;

1. INTRODUCTION
This project is based on the idea that speech and music are closely related and probably share evolutionary origins, an idea that has been explored from several perspectives in recent decades in a growing literature on the evolution and function of art and ritual [1], [2]. Based on how infants seem to learn the melodic patterns of speech utterances long before they grasp any words [3], [4], it is reasonable to believe that some aspects of creating and experiencing music can be related to the *communicative* role of musical features in speech. Not *what* we say, but *how* we say it – how the intonation, register, tempo, rhythm, dynamics, and voice quality form a communicative layer of its own in speech. In conversation, these features have clearly pragmatic functions for signaling turn-taking, highlighting important information and interpreting intentions etc., but from a musical point of view it is interesting to see how these structures also can make recognizable and meaningful patterns *as music*. The improvised nature of everyday conversation is a particular parallel to improvised music. This was the background for a recent artistic research project, exploring from a creative musician's point of view how everyday speech can be used as a source in improvised music. Following the ideas of Bakhtin [5], a main focus was to explore the musical characteristics of different *speech genres* as used in real life conversations, such as *baby talk, argument, small talk, ritual, interrogation, public speech, pillow talk* etc. As a tool in this exploration, a software-based musical instrument system was developed for analyzing, extracting and "orchestrating" musical features of speech into musical arrangements, in real time. The focus on real life situations rather than performative speech genres meant that this exploration was based on recorded conversations, covering a wide range of social situations where the genre is an integral part of the message.

NIME'18, June 3-6, 2018, Blacksburg, Virginia, USA.

1.1 Musical context
It is nothing new to use speech to make music, from the rhetorical concepts of Antiquity used in baroque music to more recently when speech became a common theme in electroacoustic music. Composers like Paul Lansky, Trevor Wishart, Peter Ablinger and many others have based much of their music on speech and the human voice. Cathy Lane has covered contributors to this field and describes a whole range of distinctly different ways speech has been used in music [6], [7]. While many have typically used poems, public speeches and other performative speech genres, or more conceptual approaches focusing on identity and personality, site specific or historic associations, the relation between *everyday conversation* and *improvised interplay* has been less explored and was therefore interesting to pursue.

1.2 Instrument context
In the context of NIME, there is also a long history of making instruments utilizing voice and speech. Recent contributions include systems that control various kinds of voice synthesis usually by means of hand gestures, either on a modified accordion [8], with gloves [9], stylus [10], guitar [11], microtonal keyboards [12], or using motion capture to track hand gestures in space [13], [14]. Other systems have taken the opposite approach, using voice as input to control other instruments [15], [16]. The present system also use speech as input, but instead of relying on expressive performer gestures it could rather be seen as a real-time compositional tool to analyze and extract, transform and arrange layers of rhythmic, melodic, harmonic and other musical features from a given speech source. This is not unlike transcribing and scoring speech melodies on paper or processing speech and composing an electroacoustic piece, but to do this through the dialogical process of improvisation it needed to be a real time and interactive *instrument-like* system.

1.3 Linguistic background
The linguistic fields of prosody and conversation analysis have provided the background for approaching significant musical content in speech in this project. Such prosodic features include for example how stressed syllables are typically used to highlight the most important words, while *high*-pitched syllables are often used to mark *new* information [17]. In addition, there is a range of musical characteristics that is typically associated with different speech genres, such as speech rate, dynamic range, metric regularity etc. [18]. These and many other interesting prosodic and semiotic phenomena has provided methods for identifying significant features that would be interesting to use as a foundation for exploring speech musically. This linguistic background has consequently influenced design choices in the software instrument used in these explorations.

2. THE ORCHESTRA OF SPEECH

The overall artistic goal of the project was to create music that could say something about the relationship between improvised music and everyday conversations. To facilitate this, the system had to be able to cater for a wide range of possible musical ideas derived from speech, from rhythmical extrapolations, implicit harmonic fields, melodic shapes etc. amounting to an open-ended laboratory for playing with these elements. Creating new musical possibilities was part of the goal.

A thorough description of the system is outside the scope of this short paper, so only a brief account of its development will be given here. First trials just included some simple means to analyze, transform and resynthesize features from a stream of speech recordings in different ways. This setup posed some challenges regarding the role of the performer. Since the output was not directly related to performer gestures it felt more like "live-remixing" than performing, something that could have been done better by composing. One solution to this was to make the system more reactive to input. Speech recordings were then organized into corpora of analyzed segments (syllables, motifs, phrases), with Markov models describing the likelihood of transitions between the segments based on musical descriptors like mean pitch, duration, amplitude, pitch slope and tempo. Like the corpus approach used in Diemo Schwarz' *CataRT* instrument [19], this organization allowed for a much more performative and musical way of relating to large collections of recordings, navigating the descriptor space to quickly access segments with certain musical qualities. But it also allowed the system to react to live audio input, triggering similar sounding speech segments or querying the Markov models to provoke (musically) probable responses and alternative sequences from the speech corpus. In line with the artistic goals of exploring the improvisational dynamics of conversation, a piano could then be used as a dialogical partner to the system and in effect act as a musical input "controller" for triggering speech segments. The aim was not to create an automatic impro-system, but rather to introduce the unexpected response and dynamic interaction that is part of improvised interplay. At the same time, integrating expressive performer gestures as an *indirect* way of producing output also brought the system closer to a performer-paradigm.

Combined with the existing features for analysis/synthesis and multilayered orchestration, this setup was flexible enough to pursue complex new musical ideas in improvised in responses to particular qualities of different speech genres, yet simple enough to perform with. Together with working out good mapping strategies for physical hands-on control of the many functions of the system, this way of using the piano for input helped achieve a greater sense of *embodiment* and what Sidney Fels has called *control intimacy* [20] with the system.

Developing further the projects artistic themes of *speech* vs *music, voice* vs *instrument* and *recorded sound* vs *acoustic live performance*, the sound producing side of this concept was extended beyond the typical speaker setup to include an array of transducers mounted on additional acoustic instruments such as drums, cymbals and stringed instruments. Following the general metaphor of orchestration, this became the *Orchestra of Speech* performance concept. Through several studies and compositions, a musical repertoire for improvising has been developed for this concept, realized in a series of performances both solo and in ensemble settings. An interactive sound installation version has also been realized, featuring a telephone as a way for audience members to interact and communicate with this speech/music hybrid system.

3. REFERENCES

[1] E. Dissanayake, "Retrospective on homo aestheticus," *J. Can. Assoc. Curric. Stud.,* vol. 1, no. 2, 2003.

[2] S. Malloch and C. Trevarthen, Eds., *Communicative musicality: Exploring the basis of human companionship*. Oxford: Oxford University Press, 2009.

[3] A. Fernald, "Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message?," *Child Dev.,* vol. 60, no. 6, pp. 1497–1510, 1989.

[4] D. Snow and H. L. Balog, "Do children produce the melody before the words? A review of developmental intonation research," *Lingua,* vol. 112, no. 12, pp. 1025–1058, 2002.

[5] M. M. Bakhtin, "The Problem of Speech Genres," *in Speech Genres and Other Late Essays,* Austin: University of Texas Press, 1986, pp. 60–102.

[6] C. Lane, "Voices from the Past: compositional approaches to using recorded speech," *Organised Sound,* vol. 11, no. 1, pp. 3–11, 2006.

[7] C. Lane, Ed., *Playing with Words*. London: CRiSAP, 2008.

[8] P. R. Cook and C. N. Lieder, "SqueezeVox: A new controller for vocal synthesis models," *in ICMC,* 2000.

[9] S. S. Fels and G. E. Hinton, "Glove-Talk II - a neural-network interface which maps gestures to parallel formant speech synthesizer controls," *IEEE Trans. Neural Networks,* vol. 9, no. 1, pp. 205–212, 1998.

[10] S. Delalez and C. Alessandro, "Vokinesis : syllabic control points for performative singing synthesis," *in Proceedings of the International Conference on New Interfaces for Musical Expression,* 2017, pp. 198–203.

[11] M. Astrinaki, N. D'Alessandro, L. Reboursière, A. Moinet, and T. Dutoit, "MAGE 2.0: New Features and its Application in the Development of a Talking Guitar," *in Proceedings of the International Conference on New Interfaces for Musical Expression,* 2013.

[12] L. Feugère, C. D'Alessandro, B. Doval, and O. Perrotin, "Cantor Digitalis: chironomic parametric synthesis of singing," *EURASIP J. Audio, Speech, Music Process.,* vol. 2017, no. 2, Dec. 2017.

[13] G. Beller and G. Aperghis, "Gestural Control of Real-Time Concatenative Synthesis in Luna Park," *in P3S, International Workshop on Performative Speech and Singing Synthesis,* 2011.

[14] G. Beller, "The Synekine Project," *in ACM International Conference Proceeding Series,* 2014.

[15] S. Fasciani, "Voice-Controlled Interface for Digital Musical InstrumentS," *PhD Thesis,* National University of Singapore, 2014.

[16] J. Janer, "Singing-driven interfaces for sound synthesizers," *PhD Thesis,* Universitat Pompeu Fabra, 2008.

[17] A. Wennerstrom, *The Music of Everyday Speech: Prosody and Discourse analysis*. Oxford University Press, 2001.

[18] T. van Leeuwen, *Speech, Music, Sound*. London: Macmillan Press, 1999.

[19] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton, "Real-time corpus-based concatenative synthesis with catart," *in 9th Int. Conference on Digital Audio Effects (DAFx),* 2006, pp. 279–282.

[20] S. Fels, "Designing for Intimacy: Creating New Interfaces for Musical Expression," *Proc. IEEE,* vol. 92, no. 4, pp. 672–685, Apr. 2004.

4. Appendix: project links

Project website: orchestraofspeech.com
Solo performance: folk.ntnu.no/danielbu/solo-performance.mp4
Sound installation: folk.ntnu.no/danielbu/installation.mp4