Deconstructing Speech: new tools for speech manipulation

EDWARD KELLY

Centre for Creative Research into Sound Art Practice (CRiSAP), 13th Floor, Tower Block, London College of Communication, Elephant & Castle, London SE1 6SB
E-mail: morph_2016@yahoo.co.uk

My research at the London College of Communication is concerned with archives of recorded speech, what new tools need to be devised for its manipulation and how to go about this process of invention. Research into available forms of analysis of speech is discussed below with regard to two specific areas, feature vectors from linear predictive coding (LPC) analysis and hidden Markov-model-based automatic speech recognition (ASR) systems. These are discussed in order to demonstrate that whilst aspects of each may be useful in devising a system of speech-archive manipulation for artistic use. Their drawbacks and deficiencies for use in art – consequent of the reasons for their invention – necessitate the creation of tools with artistic, rather than engineering agendas in mind.

It is through the initial process of devising conceptual tools for understanding speech as sound objects that I have been confronted with issues of semiotics and semantics of the voice and of the relationship between sound and meaning in speech, and of the role of analysis in mediating existing methods of communication. This is discussed with reference to Jean-Jacques Nattiez's *Music and Discourse: Towards a Semiology of Music* (Nattiez 1987). The 'trace' – a neutral level of semiotic analysis proposed by Nattiez, far from being hypothetical as suggested by Hatten (1992: 88–98) and others, is present by analogy to many forms of mediation in modern spoken communication and the reproduction of music, and it is precisely this neutrality with regards to meaning that tools for manipulation of speech must possess, since the relationships between the sound of speech and its meaning are 'intense' (after Deleuze 1968).

1. INTRODUCTION

The research project I am presently engaged in is concerned with archives of recorded speech collected by Cathryn Lane for a project entitled *The Memory Machine*. Altogether this archive consists of 8 Gb of monophonic speech recordings, from 30–300 seconds in duration. The purpose of this project is to devise a way of accessing the archive according to its phonemic content, but according to aesthetic rather than linguistic criteria. Hence we are not so concerned with what the words actually mean, but still the system needs to be accurate since it must differentiate an 'eh' from an 'ah' sound. The system should be easy to use for non-expert users, and so the representation of whatever information is chosen to type phonemic sound-objects needs to be carefully considered, not only in terms of its ergonomic suitability, but also in terms of its neutrality with regards to the signification of meaning.

2. COMMUNICATION, SEMIOTICS AND ANALYSIS

A principle of semiotics expounded by Nattiez is that there are separate, discrete systems of signification in the production, transmission and interpretation of meaning in music. The sound itself (be it recorded or spontaneous) may be analysed separately from its making or reception, and this is what Nattiez calls a 'trace' of the work, which is concerned with the structure of the material. Whilst controversial in semiotics – critics argue that analysis can never be neutral since it presupposes meaning in the configuration of materials – this mediation of meaning by analysis has a technological analogue in Linear Predictive Coding (LPC) and related speech compression techniques. In LPC, the voice is analysed in terms of pitch, formants and voiced/unvoiced segmentation, and it is the analysis rather than the recorded speech that is transmitted, from which the meaningful phrases of speech are reconstituted. In an LPC encoded conversation of the modern GSM[1] system used to transmit speech between mobile telephones, the participants do not actually hear each other's voices but rather sound re-constituted from analyses of the voice. Although the meaning is retained and understood by participants, the actual signal of each speaker goes only as far as the individual's handset. It is a trace of the form of speech that communicates meaning between the two.

A recorded signal also has parallels with the trace, whether it is recorded as magnetic changes in iron oxides (tape), physical variations in a three-dimensional surface (vinyl) or binary-encoded PCM (digital), it is an isomorphism (literal or abstract) of sound, but there are problems with the notion that meaning itself exists apart from in its production and

[1]GSM stands for Groupe Spécial Mobile, a pan-European group set up in the 1980s to develop low-cost digital mobile telephony. The method of speech encoding used in GSM phones is a Regular Pulse Excited – Linear Predictive Coder (RPE–LPC).

Organised Sound **11**(1): 73–80 © 2006 Cambridge University Press. Printed in the United Kingdom. doi:10.1017/S1355771806000100

interpretation. Communication and artistic activity alike are mediated forms, the former in sound as speech or in writing as text, and the latter in sound, in works of art and in reproduction as recordings. Whereas meaning exists in subjective appreciation or understanding of something, information may be stored, reproduced, analysed and manipulated separately from its subjective meaning. The interpretation of information results in meaning, so that neutral semiotic analysis is non-existent at the point of intervention by the analyst, or as Hatten states:

> Since the neutral level of analyses can proceed from hypotheses, and presumably those hypotheses are developed with attention to their *potential* poietic or esthetic relevance, perhaps a better term for Nattiez's neutral level would be 'hypothetical level'. (Hatten 1992)

A neutral level of analysis of speech concerns itself with properties and quantities rather than meaning, all of which may be regarded as information. Such forms of analysis are concerned with describing what kind of sonic activity exists in a recorded or broadcast signal, since the relationship between sound and meaning in speech is complex. Analysis of speech is therefore separate from actuation of meaning, a system of quantities by which speech may be described apart from the rules of grammar and syntax, but the form of analysis dictates what interpretation opportunities are offered by it. The information sent from one mobile telephone to another cannot be defined as language, rather it is instructions for making a simile of that which conveys meaning and although communication occurs, in a more direct analysis of the process the user talks to modelled, synthesised versions of people's voices. The same is true of the Internet, where binary data are reconstructed as pictures, words, sound and colour, but exist apart from human interaction only as sequences of 0 and 1.

In computer music, meaning itself is often much less clearly defined than with speech. The forms that the trace of a recording may take are its waveform, its sonograph, its envelope and a wide variety of other representations of analyses of the waveform. Each trace of sound represents a different type of information that may be useful in discovering a different feature or characteristic of sound, so in order to choose a correct analysis method for an application it is vital to determine what kinds of information are needed, and in order to make this information useful to a composer, it is necessary to look at how it may be represented.

3. A TYPOLOGICAL FRAMEWORK FOR THE ANALYSIS OF RECORDED SPEECH

If speech is to be analysed according to content then a form of organisation is required, a typology of speech, in order for analyses to be rendered useful to the composer. Many attempts at formulating such a typology are known, the most famous of these being the International Phonetic Alphabet or IPA. The IPA, as its name suggests, is an attempt to encompass all known phonetic elements across the globe into one taxonomic structure. It is to some extent successful in fulfilling this purpose. It has been used by composers before, most notably by Trevor Wishart as a tool for notating complex vocal scores. As a prescriptive tool it is appropriate, since its approach to speech is (phon)etic (after Nattiez – of the process of originating meaning) and hence concerned with the production of speech rather than its perception. Herein lies its weakness as a tool to *describe* recorded speech, and this is reflected in its terminology that is largely physiological (dental, nasal, etc.).

With regards to the perception of phonemes – the sound of speech rather than its method – it is useful to consider typologies that have been successful in describing other phenomena of sound. Pierre Schaeffer's *Tableau Récapitulatif de la Typologie* or TARTYP (Schaeffer 1966: 459) as shown in figure 1, along with the *Traité des Objets Musicaux* (*ibid.*) and other works by Schaeffer and more recently Michel Chion (Chion 1983), point to a system that has been in use as much as it has been continuously developed for over forty years. The point about the TARTYP is that it asserts the primacy of perception, so that its categories are related to the way sound is heard and understood in terms of objects, rather than its means of production. Several features of this system are apparent that are useful in describing speech sound objects. Duration is one, where the 'impulse' is distinct from the 'note', for example. Some terms are more

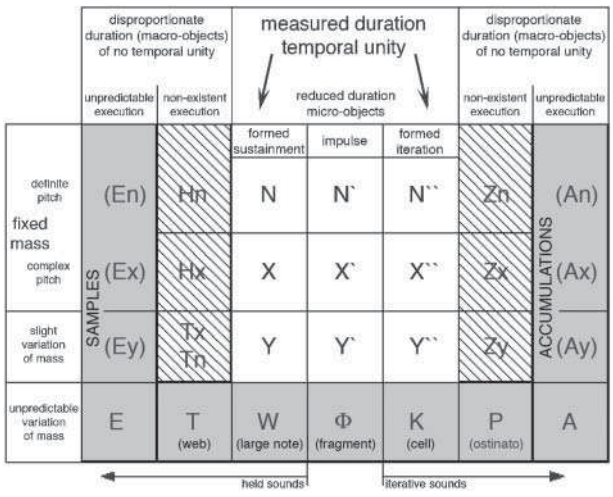

Figure 1. The *Tableau Récapitulatif de la Typologie* or TARTYP, a system of organisation for sound objects based on perceptual criteria devised by Pierre Schaeffer.

akin to methods for organisation of material though. The notions of 'sustainment' and 'iteration' both describe fragments of normal speech, if one is a second of 'um' and the other is three seconds of 'umumumumum'. The subjectivity of Schaeffer's system is manifest in its categorisations of sound; it is a system of organisation, based in an experiential philosophy motivated by a need to rationalise sound's organisation. It is left up to the individual to decide whether to take one second of sound or three, and which to define as the 'sound object'. The TARTYP is a conceptual tool as much as it is a descriptive catalogue. Rather than distinguishing between subtle differences in the pronunciation of a 'n', it is an attempt to establish a poietic framework of sound objects from perceived qualities.

Problems in using the TARTYP to organise speech are evident from the isolation of a single word. The word is a perceptual object, yet there is only one place in the chart it can occupy according to its morphology, the cell (K), although elements of the word may be placed in other categories. In fact there are no webs, few ostinati and no accumulations in natural language, although such objects may be contrived by composers. Clearly the TARTYP is a useful tool for general sound classification in the context of composition; however, it is too general a system for the classification of speech. In speech there are absolutes of perception of sound in the sense that a speaker of a particular language will recognise them as such (e.g. English vowel sounds), but their context as elements of communication define their meaning, so that in language the signifier and the signified are related by a complex web of interpretants (Peirce 1931–1935). The trace of recorded speech is divisible into its phonemic components, quantifiable unambiguous entities, whereas the processes of speaking and cognition are complex and separate. The relationship between the two processes has what Deleuze calls 'intensity' (Deleuze 1968) in that there are ambiguous and complex relationships between the elements of each process (the phonetic origination of sound and the phonemic perception of it).

It is of course possible to extract sound objects from speech that fit well into Schaeffer's system of organisation. When devising typology of speech, there are aspects of the TARTYP that are worth retaining or modifying. The notion of 'execution' may be subverted to differentiate deliberate, spoken phrases from involuntary pause sounds, breath sounds and tongue sounds. Although subtly different from Schaeffer's definition of execution, it is useful to make the distinction between sounds that are made assertively (spoken words), deferentially (pause words, 'er', 'um', etc.), and sounds that are accidental (e.g. palate–tongue interactions) or incidental (breath sounds). Within the limits of execution it is possible to define phones, syllables and words, whereas outside those limits (the grey area of the chart) are the accidents/incidents of speech. This is a problematic definition in terms of its implication that meaning is inherently part of the structure of sound, whereas the relationship between the two is not straightforward. The notion of 'sustainment' is as applicable to certain phones in assertive speech as it is to pause or deference sounds. To deconstruct the trace of speech with regard to sound alone implies that the semantic web (after Nattiez) is exclusive from the model we are trying to construct, yet the implication of 'meaningful' *vs* 'meaningless' implies a semantic differentiation apart from the sound objects themselves. Whether deference sounds have peculiarities of sonority, however, is something that may be investigated, so for the present time these are considered useful categories to implement.

The idea that duration plays a large part in our perception of sound is important and may be objectively quantified. For example, instantaneous spectra taken from 't', 's' and 'd' may be very similar in terms of their noise content and high-frequency content, yet we perceive them as distinct entities. It becomes clear from quantitative analysis of speech that the perception of speech is more naturally sophisticated than perception of sound in general, and that if a language is understood then the human ear and brain are much more finely attuned to subtleties within speech than they are to sound in general, but also to generalised classes of sound-to-language interpretation. Another consequence of such analysis is that it highlights the separation between the way semantics operates and the way sound is perceived. This is why the concept of the 'word' does not fit so neatly into Schaeffer's categorisation system, and why it is important when considering speech as sound objects to exclude semantic definitions from the mode of organisation, since the rules that govern language are separate from those that may be used to construct a 'language' of sound. The idea of deference sounds strays from this notion somewhat.

A typology of speech-sound objects needs to encompass all the features one would expect to find in a recording of normal speech. A pseudo-Schaefferian approach to phonics is embodied in the *Table of Phonemic Morphology* (figure 2).

4. CONTENT-BASED SEGMENTATION OF SPEECH

Whilst there are tools available for the construction of a trace of the voice (e.g. LPC, Mel-frequency cepstral[2]

[2]A cepstrum (pronounced 'kepstrum') is the result of taking the Fourier transform of the decibel spectrum as if it were a signal. This resolves periodicities in the spectrum and so is often used to deconstruct sound to reveal basic pitch and formant information.

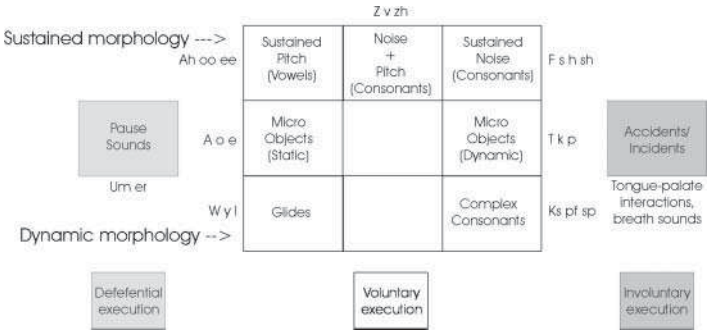

Figure 2. Attempts to categorise phonic classes according to sonic properties, as well as *execution* of vocal sounds are embodied in the *Table of Phonemic Morphology*.

analysis), they are generally found to be a means to an end separate from the source itself, although their assimilation into tools for the composer may be fruitful. Two types of voice-specific system are discussed here, both of which were developed for reasons other than sound manipulation.

LPC and GSM arose out of a need to compress speech so that more channels may be transmitted down a telephone line or radio signal, and it is economic rather than artistic pressures that have shaped their emergence. The development of technology for the recognition of speech by computers is an area in which engineering challenges have been pursued for partly economic, partly political reasons. Automatic Speech Recognition (ASR) systems were developed initially by DARPA (a research foundation set up by the American military in 1958) in order that fighter pilots could control aeroplanes by voice alone. These are examples of tools that reflect the motivation behind their design and may or may not fulfil that purpose for which they are designed. Although the applications for which existing systems were devised may be of less interest to the composer, it is worth investigating such systems in order that useful techniques employed within them may be borrowed for artistic applications.

ASR systems perform a type of categorisation of recordings and so it would appear that they offer useful means of discerning one type of sound from another. The representation of sound in a Markov/ASR system is a cepstral transform based on the Mel scale of pitch, a perceptual measure based on experimental scientific inquiry. The nature of such analysis is that each state represents steady states of pitch and formant information. The only information provided about the morphology of speech is the probability of transitions between states. An element of speech is dynamic and changes happen with different degrees of smoothness, so packages like the Hidden Markov Toolkit[3] introduce the concept of the tied phone.

[3]http://htk.eng.cam.ac.uk/

Since ASR systems are designed to map speech patterns onto the web of semantics that makes up written rather than spoken language, the results of ASR systems can be erratic. ASR systems attempt to model logically a process with Deleuzian intensity, and one in which semiotic relationships are inverted; the signifier (speech) becomes the signified in a relationship whose complexity means that the text can only partially refer to the speech.

The form of linear predictive coding used in modern European mobile telephony (Regular Pulse Excited – Linear Predictive Coding) splits the signal into a set of coefficients for reconstruction of the formant peaks by filtration, and a codebook of voiced, unvoiced and transitory excitation impulses. Information about the excitation signal along with the filter coefficients is transmitted in RPE–LPC and this is used to reconstruct the speech at the other end. By using a fixed codebook of excitations it is possible to reconstruct realistic sounding speech at the receiver. Although this codebook contains discrete states (20 ms frames of information) organised arbitrarily, it introduces the concept of a codebook as a means of identifying a particular sound with a position in an organised matrix of possibilities. This is not a form of segmentation that bridges the gap between the mechanism and the perceptual object; the elements of the codebook do not represent what are perceived as phones, but the concept can be extended further to demonstrate how content based speech segmentation systems may be constructed that deal with recorded sound directly, rather than the process of transmission of meaning in the form of a model.

All the systems looked at thus far model speech as a series of discrete states initially. Both systems chop the material up into much smaller sections than are perceived as phones by the listener, and whilst one attempts to construct text from discrete cepstral frames mapped to a statistical model, the other sequences micro-sound objects through a filter in order than speech may be reconstructed, but without any attempt

to construct higher-level macro-units of speech (phones) that to the listener are the smallest meaningful part of speech. Analysis techniques have concentrated on much smaller fragments or steady states only, and so there is a gap between what is defined as a unit in machine terms and in human terms. Markov models are an attempt to bridge this gap, with the notion of tied phones, but most ASR systems are too specifically trained for use in generalised segmentation of a large archive. Where an ASR system begins with painstakingly specific annotated speech recordings and builds a picture of the general traits of a person's language, often with higher-level language parsing built into the model as linkages between entries in a textual database,[4] it is defined from the outset by the language, accent and dialect with which it will be trained and by which it will be fixed in its use. A compositional tool on the other hand has to be open ended, flexible in its use and non-deterministic of its outcomes. The sonic artist's prerogative is to execute the decisions by which a work is made, so to contradict that imperative by limiting the tools of the composer to one language, one dialect or one voice is to render ineffectual the tools themselves.

ASR errors demonstrate how different the characteristics of the human voice may be from one person to another, and so generalised tools to deal with speech need localised models, where each set of analyses may be compared with others but is distinct from them. What is needed are tools for manipulating a 'codebook' of phonemic utterances that learns on-the-fly, so that a database of classes may be built up independently of a language model. Intervention by the user may take the form of setting identities for each class after a single pass based on generalised phonemic classes. Errors in classification may also be pointed out, so that a second pass of the classifier may refine its accuracy. The one thing lacking with regard to speech processing software, the focus of my research at LCC, is a self-calibrating phonemic classification and manipulation system.

Far from being a rigidly defined system, where input material affects the usefulness of such a tool according to the limitations in software, a system for artistic manipulation of speech should involve a high degree of interaction based on the 'primacy of the ear' (Schaeffer 1966), and yet in order that such manipulations may be related to content, a robust but flexible modelling system is required. The unpredictability of possible input material with regards to dialect, accent and language further emphasises the need for flexibility, even extensibility of an interface to recorded language, as is made clear by expanding on a theme of semantics manifest in the differences between speech cognition in speakers of different races.

According to the International Phonetic Association, there are about 112 distinct phonetic sounds produced in human speech. Whilst this figure may be contentious,[5] it shows how vast the range of sounds of speech is. What it does not do is provide insight into the relationship between phonetics and phonemics, and how conditioning affects the cognition of language and categorisation of its sounds to a large extent. For example, in Hindi there are various ways to pronounce 'n', three of which are according to the IPA, the palatal nasal ɲ, the alveolar nasal n, and the retroflex nasal ɳ. To the vast majority of Westerners these three sounds are heard as the same linguistic entity, and many are unlikely to differentiate between them. To an Indian, however, they are three distinct phones that are used in different contexts. Further east in China there is little differentiation between a 'r' sound and a 'l' sound, and many languages do not have a 'w' at all. Far from being a linguistic cliché, this is a common occurrence between languages, so that one phonemic perception of language is different from another.

Tools for the manipulation of speech should therefore have extensible models of speech. Research into ASR systems shows that inflexible databases of statistics about one set of material are only robust in terms of that material, and when applied to other corpuses of speech their accuracy and usefulness are both challenged. An artistic application for manipulating speech requires no assumptions to be made in terms of the source of material. Manipulations of sound are neither 'right' nor 'wrong' in terms of aesthetic merit, but judged 'appropriate' or 'inappropriate' according to the composer's prerogative. Such tools must therefore begin with unfixed assumptions and facilitate the decisions of the user, but with an underlying sophistication so that morphologies may be classified.

It is because the semiotic 'text' of recorded speech is much more complex than just the syntactical organisation that it is necessary to develop new tools for composition with language. In many senses it is more accurate to refer to the 'story' of a given recording, by which it is inferred that there are origins, context and purpose, rather than just dry syntax

[4]Such as the online referencing system Wordnet (http://wordnet.princeton.edu/) organised around current psycholinguistic theories of memory.

[5]There are many more click sounds used in sub-Saharan Africa than are shown on the International Phonetic Alphabet. This is just one example of how typologies of language generalise, and whilst the IPA is a schematised map of phonetic possibilities, its complexity demonstrates how problematic it is to construct a global taxonomy of speech.

and grammar. Aspects of the story are contained within perceptions of difference or of empathy, clear examples of which being gender differentiation and the perception of regional accents. Finally, the prosodic aspects of speech – intonation, rhythm, pitch and stress – carry aspects of the story regarding the individual, their origins, emotional state and personality, as well as elements of emphasis and expression. These are traits of language likely to be attractive to composers working with speech, so it is vital that software tools are developed that facilitate interaction with these aspects of speech.

Such tools do not need to be developed from scratch necessarily. Some of the techniques discussed earlier may be co-opted. LPC analysis is a particularly useful method for discovering the fine structure of formant patterns. Another mathematical technique – the average magnitude distance function (AMDF) – can be used to determine voiced from unvoiced speech. Since meaning in speech is contained in so many aspects of its delivery, the logical approach to developing tools for its manipulation is to attempt to deconstruct the voice into multiple feature-vectors so that models of specific spoken elements may be objectively described by parameter sets. Specific sound objects may then be identified by the user and used as a reference for similar objects by storing their parametric representations in a database. Furthermore, such parametric representations of speech may be used in part to reconstruct (re-synthesise) elements of speech with different excitations, pitch curves and rhythms. The most important principle of such a system is that the primacy of the ear is preserved, so that the user makes decisions as to how an object is delimited and identified. It is the composer's interpretation of the material that influences the structure of the database, rather than some pre-defined inflexible system such as the IPA.

5. THE TREE APPROACH

A pragmatic approach to segmentation is to determine what is the first distinction that may be implemented in some way. The voicing of a particular sound is one possible distinction, after which further distinctions may be made according to easily distinguishable feature-differences between two groups of speech-elements. With unvoiced sounds this may be the attack-time, with voiced it may be the noise content, or formant positions. A user fluent in Hindi may well make the decision that ŋ and ˣ are separate sound objects, but since this distinction is made entirely by the user in the context of a customised categorisation process, it represents just one possibility for use, and is not exclusive of other taxonomies – its contingencies are defined with respect to the material by the user's application of his or her own experience.

6. AESTHETICS AND AMBIGUITY

The criterion on which both LPC and ASR are judged is intelligibility. If the speech at the receiver end of an LPC encoded conversation is recognisable and understood, then the technology is successful. Likewise, if the text of a transcript of a radio programme is grammatically and syntactically accurate, then the ASR system is working, well!

But Aesthetic criteria are ambiguous, and perhaps do not need such clear distinctions or such accurate models. A relativistic, subjective model such as Schaeffer's may be artistically useful without being deterministically 'right', and by specifying less specific detail about the categories there is more ambiguity in such a system. Schaeffer's categories serve to delineate potential juxtapositions or areas of inquiry. Perhaps in considering speech for artistic purposes there is no need for linguistic criteria of 'vowel', 'consonant', etc., and a more quantitative approach should be taken?

Research into speech software has until now either been focused on engineering solutions to military or commercial problems, on medical applications for the treatment of speech defects, or on the academic study of linguistics. The latter category of software for linguistic analysis (e.g. Praat,[6] Wavesurfer[7]) involves a great many analysis techniques, but these are presented in order that the linguist may manually annotate recordings of speech and form conclusions about language. Just as the text is the focus of ASR systems, the theories of language and conclusions of the linguist are the focus of linguistic analysis software. A pattern is emergent that the function defines the form of such software. The gestures of speech may be analysed phonically and according to pitch and rhythm, but these aspects need to be made available to the composer in a way that is relevant not only to phonemic but also to musical concerns. A shorthand script for pitch gestures and a scansion-like analysis of rhythm are more accessible to the composer than a graph of frequencies and a table of durations, so interfaces need to be designed that reflect the concerns of composers that may not necessarily represent analysis data in a rigorous, scientific way but rather by employing an intuitive graphical scheme. Fortunately the analysis techniques employed by Wavesurfer are available as a scripting language (Snack[8]) in which applications may be coded. This, along with other frameworks currently available (particularly Marsyas [Tzanetakis and Cook 2000] and Clam [Amatriain, Arumi and Ramirez 2002]) are making development of such bespoke applications much more practical than ever before.

[6]http://www.fon.hum.uva.nl/praat/
[7]http://www.speech.kth.se/wavesurfer/
[8]http://www.speech.kth.se/snack/

7. WORK IN PROGRESS

Initially, the long-term aim was to devise such a system in the form of a stand-alone application encompassing all the principles of open-ended design, parametric deconstruction of material, and tools for manipulation of material according to user-defined datasets. A more immediate goal was the construction of a prototype system from discrete classes of process in a music-based environment such as Max/MSP[9] or Pure Data[10] (PD). Collaboration between the author and Nicolas Chetry at Queen Mary University of London has produced LPC and AMDF external objects for Pure Data (see figure 3). One thing has become clear through this process of prototyping externals for such functions (both are conventional math routines but had yet to be developed for PD) – that although such analyses may be possible in real time, it is perhaps not the best way to approach the problem.

The eventual goal of our project is to create software that constructs a database of analyses, on which basis the soundfiles themselves may be manipulated. Such a database may be time-consuming to create, but its requirements for use may not be restricted to an offline-segmentation or manipulation of the sound. Certain compositional possibilities are raised by the existence of some externals – the lpc~ external may be used to set up a real-time vocoder, for instance, but if the archive is a pre-existent asset then it is unnecessary to calculate AMDF functions or LPC coefficients in real time, since any SQL database made offline is accessible from within PD via Ian Mott's sqlsingle object.[11]

8. NEW AND EXISTING FEATURE VECTORS

The nature of my research project necessitates the generation of large amounts of data. The ability to perform many different types of analysis results in many different types of comparison that may be made, and also different potential uses of the database in performance, manipulation and comparison between phonic classes. It is worth examining an ensemble of feature vectors to maximise the creative potential of the system. As well as cepstral coefficients, line-spectrum pairs from LPC analysis, and voicing and pitch analyses, other less conventional features vectors are being examined as to their specificity to different phonic classes, such as the highest significant spectral component (figures 4 and 5).

Such tools may be useful, but without interpretation of the data in software they are no more use than standard linguistic analysis software. It is in the second phase of this research, when a database is available of feature vectors, that open-ended strategies for classification may be implemented. Recent advances in artificial intelligence point to ways to do this, not least the work by Elias Pampalk at the Austrian Centre of Artificial Intelligence (OFAI), whose drum classification engine suggests an intuitive interface and categorisation system[12] concerned not with boundaries but with a representation of timbral similarity.

9. CONCLUSION

Speech classification systems have been available for a number of years, but their application to artistic practice has been greatly overlooked. Our research here at LCC aims to remedy that, to deliver an application that greatly expands the creative possibilities for artists working with the human voice. Such tools as are available for manipulation of speech have been devised for non-artistic reasons, and their form fits their function making them largely unsuitable for artistic uses. The creation of a piece of software that is

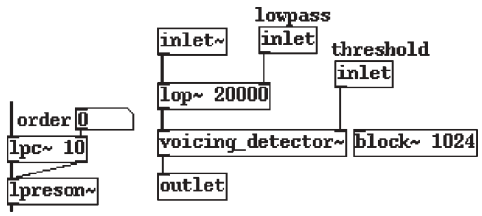

Figure 3. Speech processing tools such as lpc~ and the amdf voicing detector are quick to prototype by writing C classes for PD.

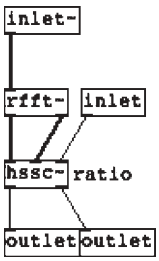

Figure 4. The hssc~ object for Pure Data detects the highest significant spectral component, thus Hssc = fmax, where $\alpha > (\alpha\text{max}/\text{ratio})$.

[9]http://www.cycling74.com
[10]http://puredata.org
[11]http://www.reverberant.com/PD/index.htm

[12]http://www.ofai.at/~elias.pampalk/dafx04/. The IPA font used is SILDoulosIPA-Regular available from http://scripts.sil.org/. PD externals are available from the Pure Data External Repository at http://pure-data.sourceforge.net/

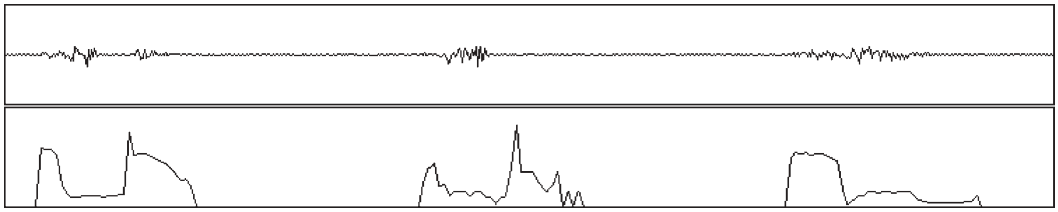

Figure 5. The HSSC trace shown on the lower graph for the words 'cat', 'pig' and 'snail', with the ratio value set to 12. The sharp peaks at the 't' of 'cat' and the 'g' of 'pig', the high, flat profile of the 's', the stunted and elongated peaks of 'c' and 'p', and dips in the profile, a short dip for 'n' and a long dip for 'l', are all useful features in determining what kind of sound is present in the speech.

open ended, yet sophisticated enough to discriminate between the subtle nuances of human speech, is a formidable challenge, but the prospect that in the future we may access archives according to their content makes this a truly exciting piece of research to be involved in.

REFERENCES

Chion, M. 1983. *Guide des objets sonores: Pierre Schaeffer et la recherche musicale*. Paris: INA-GRM/Buchet-Chastel.

Deleuze, G. 1968. *Difference and Repetition*, trans. P. Patton, 1994. New York: Columbia University Press.

Hatten, R. S. 1992. *Music Theory Spectrum* **14**(1): 88–98. Los Angeles: University of California Press.

Ladefoged, P. 1962, 1996. *Elements of Acoustic Phonetics*, 2nd edn. Chicago: University of Chicago.

Nattiez, J.-J. 1987. *Music and Discourse: Towards a Semiology of Music*, trans. C. Abbate, 1990. Princeton: Princeton University Press.

Peirce, C. S. 1931–1935. In L. Hartshorne and R. Weiss (eds.) *Collected Papers, Vols. 1–6*. Cambridge: Harvard University Press.

Schaeffer, P. 1966. *Traité des Objets Musicaux*. Paris: Editions du Seuil.

Tzanetakis, G., and Cook, P. 2000. Marsyas: a framework for audio analysis. In *Organised Sound* **4**(3).