Speech re-synthesis with harmonic frequencies for increasing noise robustness in ASR systems

M.A. Chesnaye, in collaboration with M. R. Burghardt. **Supervisor;** B. Valkenier

September 5, 2011

Abstract

This article promotes the use of harmonic frequencies for automatic speech recognition (ASR) systems by demonstrating both their robustness in noisy conditions as their utility in speech segregation techniques. It furthermore provides a brief review of the high-level, functional organisation in the human auditory system, which is found to be reflected by ASR research, albeit at a conceptual level.

1 Introduction

Neuroscientific and psychoacoustic evidence have provided us with a high-level understanding of the signal processing components in the human auditory system. These components enable the human auditory system to resolve speech in a wide range of environments; something that ASR systems are currently lacking. Defining the signal processing components in the human auditory system is therefore important, as they can help to inspire artificial solutions.

The human auditory system can be decomposed into four general functionalities; (i) front-end signal modification, (ii) a form of sound localization and segregation, (iii) some way of imposing and exploiting syntactic and semantic restraints, and (iiii) a neural crux for integrating knowlegde and linking frequencies with semi-discrete brain regions (corresponding, perhaps, to words or phonemes).

Progress in neuroscientific and psychoacoustic research has not gone unnoticed by the ASR community. Indeed, it has been found to be focused on roughly the same functionalities as those mentioned above, that is; (i) preprocessing sound segregation techniques for differentiating the input into independent auditory streams, (ii) the exploitation of syntactic and semantic restraints for reducing the total number of possible matches, and (iii) an auditory crux, both for integrating different kinds of knowledge as for determining the label of some collection of frequencies.

Succesfuly constructing and integrating these components into a single, complex system should therefore resolve a number of problems currently encountered by ASR systems. These have been defined by Juang and Rabiner (2004) as: (i) handling speech disfluency, (ii) handling noise distortions, and (iii) a means of providing some form of feedback.

This article addresses the second challenge,

noise, by presenting a novel set of noise robust features; the harmonic complex frequencies. These are extracted during an auditory grouping procedure, after which they are used to reconstruct the original signal. The method thus not only segregates the input into meaningful streams, but also transforms them into noise robust representations.

2 The human auditory system

This section tracks a signal's ascending pathway through the auditory system. The path originates at the ear (2.1), which, after a process of selective frequency amplification, transforms the sound wave into electrical pulses. The signal is then transmitted through the brainstem (2.2) and the midbrain (2.3), into several signal processing regions of the cerebral cortex (2.4).

2.1 The ears

At the lowest signal processing level lies the ear, which transforms sound waves, composed of compressed and rarefacted air, into electrical signals. The ears play an essential role in sound localization, for which the auditory system relies almost entirely on ear-induced spectral cues (Butler & Belendiuk, 1977).

Starting at the pinnae, entering sound waves ranging between 1.5 to 7 khz are amplified by a factor that depends on (i) the structure of the pinna and (ii) the angle on the sound's source. This relative frequency amplification functions as an approximate sound localization device by means of spatial frequency differentiation in the pinnae, which is thus determined by the pinna's structure. It has been suggested that identifying these relatively stronger 1.5 to 7 kHz frequencies is equivalent to approximating the sound's source on the vertical plane (Hirsch, 1950).

Frequencies are further modified in the auditory canal through a process of constructive and destructive interference (Békésy, 1941), after which they are transformed into mechanical energy by the tympanic membrane and a series of connected bones called the malleus, the incus and the stapes. The stapes then disperse the energy onto a fibre called the basilar membrane, which causes vibrations and the displacement of very thin and sensitive hairs. These displacements then stimulate the hair's cell so as to produce a stream of electrical pulses that are transmitted for further processing through the auditory nerve.

The basilar membrane is an important signal processing component, which has received a significant research effort. Back in 1932, Békésy observed the spatial positioning of the vibrations on the basilar membrane to be responsible for the discrimination of different frequencies. Békésy proposed a linear relation, concerning frequencies and their position on the basilar membrane, which was later amended to be, in fact, logarithmic (Stevens & Volkmann, 1940).

This tonotopic, nonlinear organisation is reflected in both the anatomy as the functionality of ascending auditory brain regions, in particular: the cochlear nucleus, the superior olivary complex, the dorsal nucleus of the lateral lemniscus, the central nucleus of the inferior colliculus, and the primary auditory cortex (Cheveign, 2001). Furthermore, it accounts for a variety of properties found in, amongst other things, the auditory nerve (Johnstone, Patuzzi, & Yates, 1986), and it also clarifies our (enhanced) competence at frequency differentiation in the 1 to 5 kHz range.

A final signal processing element concerning the ears is due to their spatial positioning. Strutt

(1907) uses this to define a pair of spectral cues - again, utilized for sound localization - coined the interaural time difference (IIT) and the interaural intensity difference (IID). The first is due to the fact that when the head is turned away from the source, sound-registration in the two ears takes place at two slightly different points in time. The second, the IID, is also based on the heads angle on the source, that is, the head can block a portion of the signal's energy (called head shadowing), thus resulting in a slight amplitude difference between the two ears.

2.2 The brainstem

The brainstem and the midbrain are two of the oldest parts of the brain. Evolutionary evidence indicates that they developed through the need for navigational control in nocturnal environments (Stebbins, 1980), and are therefore mainly involved in sound localization. Nonetheless, sound localization has been found to have a significant influence on speech recognition performance (Hirsch, 1950).

The gap between the ears and the brainstem is bridged by the auditory nerve. The auditory nerve's main function is to tonotopically encode and relay inner ear induced information. It is also engaged in signal modification by means of spontaneous neuron activity, which introduces a small amount of noise to the signal.

The auditory nerve terminates in the dorsal and ventral cochlear nucleus; collectively known as the cochlear nuclei (CN). The CN fulfil two signal processing tasks: sound localization and a form of frequency amplification.

Starting with sound localization, the CN function as a waystation for integrating ear induced spectral cues with other sensory modalities, such as the somatosensory information on the position of the ears (May, 2000). The relevant spectral cues concern the frequencies that were amplified by the pinnae and the auditory canal, which presumably provide localization estimations for the vertical plane.

The CN's second function, frequency amplification, is based on the inhibition of broadband sound stimuli so as to make the narrowband stimuli more salient (Oertel & Young, 2004). This is accomplished by, in a nutshell, the separation of broadband from narrowband stimuli by two groups of cells: the tuberculoventral cells, excited by narrowband stimuli (Spirou, Davis, Nelken, & Young, 1999); and the D-stellate cells, excited by broadband stimuli (Winter & Palmer, 1995). When excited, the tubervuloventral cells inhibit narrowband sound representations (which are represented by yet another group of cells); however, when D-stellate cells are excited, they in turn inhibit the tuberculoventral cells, thus resulting in more salient narrowband stimuli representations. The CN are hereby able to detect peaks in the sound spectra, thus functioning as approximate feature extractors for (broadband) noisy environments (Oertel & Young, 2004).

A second auditory brainstem structure consists of a collection of brainstem nuclei called the superior olivary complex (SOC). One of the SOC's main functions in speech recognition is, again, sound localization. This has been empirically determined during the study of a subject with a lesion to the SOC (Griffiths et al., 1997). The subject had no difficulty in discriminating different amplitudes or frequencies, but was unable to determine the direction of a sound's source.

The SOC's essential role in sound localization is due to its binaural sound integration; it is the first place where information ascending from both the left as the right ear coincides. The SOC

uses ultra time sensitive cells and temporal information preserving mechanisms in order to capture and compare localization spectral cues from the horizontal plane (Oertel, 1999). Ultimately, its goal is to estimate the IIT and the IID.
The CN's tonotopic organisation is also preserved in the SOC, with higher frequencies residing in the lateral olive (Guinan, Norris, & Guinan, 1972) and lower frequencies in the medial olive, which has the further property of being larger, thus reflecting our bias towards lower frequencies (Warr, 1982).

2.3 The midbrain

The first midbrain structure receiving innervation from the brainstem is the lateral lemniscus (LL). The signal processing functions of the LL are obscure, yet it has been found to take an important, downstream, inhibitory role in terms of temporal information processing (Saint Marie, Schneiderman, & Stanforth, 1997), and is thus indirectly related to sound localization. This is supported by Ito et al. (1996), who observed how a LL lesion preceded sound localization impairments on the horizontal plane.
The auditory LL axons terminate in the inferior colliculus (IC), which consists of three nuclei, collectively containing more neurons than all other subcortical auditory nuclei combined (Kulesza, Vinuela, Saldana, & Berrebi, 2002). The IC is known for its high connectivity amongst other sensory brain regions, such as the superior colliculus (vision), and the paralemniscal regions that play a role in pinnae movement (Mei, 2009). This suggests that the IC is capable of exploiting certain visual and somatosensory features, which enable it to filter self-affected sounds, such as chewing (Shore, 2005).
Furthermore, the IC responds nonlinearly to changing IITs, suggesting a feedback loop to, amongst other things, the motor system for readjusting the head, the shoulders and the pinnae (Spitzer & Semple, 1993).
IC activation projects to the medial geniculate nucleus (MGN). Surprisingly, the MGN receives innervation from a large variety of senses, including tactile (touch), photic (light), caloric (heat), nociceptive (pain) and, of course, sound, thus forming a major sensory integrating waystation (Wepsic, 1966). How, or if, the auditory system exploits these features is unknown.
The MBN furthermore appears to have several tonotopic, possibly overlapping, frequency representations, which relay information to other large brain regions (Calford & Aitkin, 1983). It is also the first auditory brain region that projects to the amygdala (LeDoux, Farb, & Ruggiero, 1990), thus resulting in the production of emotional memories.

2.4 The cerebral cortex

Located in the temporal lobe lies the neural crux of the auditory system; the auditory cortex (AC). The AC has recently been decomposed into three functionally discreet regions called the core, the belt and the parabelt (Chevillet, Riesenhuber, & Rauschecker, 2011). This decomposition is based on the amount of spectral complexity needed before neural activation can occur.
At the most elementary level lie the tonotopically organized core neurons, which are activated by simple, pure tones. This suggests the registration of the most basic signal properties, such as the pitch and the amplitude. At the next level, responding to band-passed noise, lie the belt neurons, and at the most complex level lie theparabelt neurons, responding to spectral

complexities similar to that of a vowel.
Neural activation in the auditory cortex has also been measured for visual stimuli, in particular for lip-reading, and has furthermore been found to be strongly regulated by attention (Calvert et al., 1997), which is, in turn, involved in higher level feature processing (Petkov, Xiaojian, Alho, Bertrand, Yund, & Woods, 2004) and lower level, auditory brain region regulation (Suga & Ma, 2003).
A final auditory brain region that should be mentioned is Broca's area. Allthough Broca's area is more commonly known for its speech production functionalities, it has also been found to be involved in speech recognition by means of an artificial grammar (Bahlmanna, Schubotza, & Friederici, 2008).

Summary
The brief review of the human auditory system given above depicts a system consisting of four high level functionalities: (i) feature extraction and signal modification by the ears, (ii) sound segregation and localization in both the brainstem and as the midbrain, (iii) differentiation and registration of certain signal properties, such as the pitch and the timbre in the auditory cortex, and (iiii) the exploitation of syntactic restraints by means of an artificial grammar in Broca's area.

3 ASR systems

ASR systems use machine learning and pattern recognition to link a set of acoustic features with some meaningful label. Research during the last 50 years has provided the community with a number of techniques for constructing such systems. The first, dating back to the 60s, were time-normalisation methods, dynamic programming techniques and filter bank analyses. The 70s then brought linear prediction and clustering algorithms, and the 80s brought the more statistically oriented neural networks and hidden Markov models (HMMs). For a more exhaustive recitation, see Juang & Rabiner (2004).
HMMs have their roots in Bayes decision theory (Theodoridis & Koutroumbas, 2003), and are capable of representing an arbitrary collection of speech components with a set of parameters. The parameters are iteratively approximated with the Baum Welch algorithm (Baum, 1972), which shifts the current set towards those parameters optimally representing the utterance in question. Given a sufficiently large training database, HMMs are hereby capable of implicitly modelling the variability found in speech.
The techniques mentioned above form the backbone of an ASR system. Taking a slightly simplistic perspective, one might say that these techniques strive for a functionality equivalent to that of the auditory cortex, that is to say, matching collections of features with labels, or from a neural perspective - the excitation of semi discreet brain regions (responding, perhaps, to words or phonemes) by a collection of action potentials, chemicals, etc. The lower auditory brain regions, focused mainly on sound localization, are missing, as are certain lower level preprocessing filters, a context imposing syntactic and semantic restraints, and the exploitation of other sensory cues. The last is obviously not essential, since blind people, for example, suffer no speech recognition impairments. The integration with these sensory systems could therefore be omitted without endangering the possibility of an optimal recognition performance.
These bare ASR systems are capable of delivering high recognition rates under optimal condi-

tions, that is to say, in an environment that is (as close to) identical to the environment where the models were trained. However, contrary to human speech recognition, ASR performance plummets in suboptimal conditions, which suggests that something is missing. The bare ASR system should therefore be augmented, or modified, with new functionalities.

These innovations and modifications strive to solve a number of persistent problems, identified by Juang & Rabiner (2004) as: (1) handling speech disfluency, (2) handling noise distortions, and (3) the necessity of some form of feedback.

1) The first is a summation of a variety of properties encountered during natural speech production, such as the use of 'out of vocabulary' words, silences and stutters, non-grammatical constructs, and ill-formed sentences, of which the last two apply exclusively for ASR systems augmented with syntactic or semantic restraints.

2) The second challenge is perhaps the most prominent as it is not always possible to circumvent. Noise, defined simply as all the unwanted frequencies, results from environmental stimuli that distort the acoustic representation. This distortion increases the gap between the learned representation and the utterance to be recognized, thus decreasing ASR performance.

Noise is usually categorized as constant or random, and can originate from an almost infinite amount of sources. Nevertheless, they have been organised by Bellegarda (1997) into three categories: (i) variations in the hardware, such as the microphone and the available bandwidth, (ii) all externally induced and unwanted frequencies, including reflection and reverberation effects, and (iii) speaker dependant and speaker independent articulatory variations, including the Lombard effect (where speech intensity is increased according to the amount of environmental noise present).

3) The third problem is based on the fact that human communication is a two-way process, usually expressed in the form of query and answer, but also as simple as a confused or enlightened facial expression. Bare ASR systems have no means of establishing such complex feedback loops. However, the fact that people have no trouble with speech recognition whilst listening to the radio or the television suggests that such feedback loops are not essential. Indeed, the main advantage of ASR systems with feedback loops is their control over the semantic discourse, which can be exploited so as to reduce the number of possible in- and outputs. Although it has been shown that knowledge of the semantic discourse can indeed increase ASR performance (Mohamed, Dahl, & Hinton, 2009), feedback loops are presumably applied for the enhancement of speech interpretation (not recognition), and could therefore be omitted from an ASR system without endangering the possibility of an optimal recognition performance.

To conclude; ASR challenges for bare ASR systems - and, to a lesser degree, ASR systems in general - stress the importance of extending ASR systems with the functionalities found in the human auditory system. As shall be seen below, efforts to reconstruct and integrate these functions can indeed significantly improve ASR performance, as they moderate the severity of the challenges described above.

4 ASR augmentations

This section focuses on innovations and modifications to bare ASR systems that, if not entirely resolve, moderate the severity of ASR challenges. Research in three categories - (1) sound localiza-

tion, (2) context, and (3) noise - is (very briefly) described.

1) Most sound localization techniques use implementations with multiple microphones (microphone arrays), and have their roots in a theory given by Jeffress (1948), which is, at a conceptual level, biologically inspired by the human auditory, IIT resolving brain regions. The theory states that sound enters the system through atleast two sources (the microphones), after which it is differentiated into individual frequencies. These frequencies are then passed through a collection of receptors where minor time delays are imposed. Onsets of matching frequencies from different microphones are then compared so as to determine which frequencies match best in time. The value of the delay in time from the best match can then be used to determine the source of the sound.

Building on Jeffress's theory, successful implementations have been achieved (Takanishi, Masukawa, Mori, & Ogawa, 1993). It is, however, not the only option as others have shown, for example, how the exploitation of pinnae induced spectral cues can be sufficient (Tomoko, Toru, Makoto, Ryuichi, Ikuro, & Zenta, 2006).

Furthermore, sound localization is closely related to sound segregation, perhaps more commonly known as computational auditory scene analysis (CASA) (Wang, 2006), which is, in turn, closely related to human auditory scene analysis (ASA) (Bregman, 1990). ASA and CASA can be seen as a two staged process: (i) the extraction of a number of primitive features that are used to (ii) assemble frequencies into meaningful groups. These primitive features correspond to certain signal properties such as the onset, the fundamental frequency, and the harmonicity.

The grouping process that assembles frequencies into independent streams can take a top-down, schema-based or a bottom-up, data-driven approach. The top-down variant matches certain combinations of features with a set predefined groups (a form of pattern recognition) whereas the bottom-up approach relies exclusively on primitive features for re-assembling frequencies into independent streams. Data-driven approaches thus have the advantage of being more dynamic as they are independent of a set of pre-defined structures.

2) Moving on to context; ASR systems can exploit the syntactic and semantic restraints found in a language so as to decrease the number of possible matches. Syntactic restraints, found at the phoneme and word level, simply state which combinations of speech components are grammaticaly correct. This can be extended to include the possiblities that some grammatical construct will be uttered.

Restrictions at the phoneme level are captured by n-grams where the n states the number of neighbouring phonemes taken into account when categorising some utterance. Applying n-grams enables the models to represent co-articulation induced acoustic variations.

Restrictions at the word level are usually captured by an artificial grammar, which defines what word category (such as nouns, verbs and adjectives) combinations are possible. Grammars can be either absolute (also called logical of categorical) or probabilistic. Absolute grammars impose greater restrictions and, under certain conditions, can lead to higher recognition rates; however, as absolute grammas have trouble coping with ambiguity and speech disfluency, these conditions are sparse.

Probabilistic grammars solve these problems in a similar fashion as the parameter tuning process during HMM training, that is, by iteratively modifying a set of variables that approximate the

probability of uttering some grammatical construct.
A final ASR augmentation concerning context includes the use of knowledge at a higher, schema-driven level. Recognized utterances are hereby used to construct a rough concept of the current topic, which can then be used to infer expectations.
An ASR research area focussed on such a procedure is called keyword spotting. As the name suggests, a high priority is given to the recognition of specific words. These words function as cues (keywords) for inferring the semantic discourse, which can then be used to reduce the number of possible matches. Another advantage of such approaches is that they function under suboptimal conditions, since the correct recognition of a large number of utterances is deemed to be, in part, irrelevant.
3) The last topic to be discussed is on techniques for increasing noise robustness, for which three general approaches can be applied (Yifan Gong, 1995): modifying or extracting noise robust features, (ii) adapting the models to noisy conditions, and (iii) estimating the noises acoustic spectrum in order to remove it from the signal's representation.
Two of the most widely used front-end features are Mel-Frequency Cepstral Coefficients (MFCCs) (Davis & Mermelstein, 1980) and Perceptual Linear Prediction (PLP) coefficients (Hermansky, 1990). Although PLP features have been shown to be more robust to noise (Qifeng, Iseli, Xiaodong, & Alwan, 2001), both suffer significantly from noisy environments. Furthermore, PLPs have been shown to be biologically more plausible (Hermansky, 1990), yet preference is usually given to MFCCs due to their computational efficiency.
The second and third approach, called multi-condition training and model compensation, are implemented pre and post-training respectively. Multi-condition training involves the intentional contamination of the training material with noise. The models are then trained to represent noisy data, thus leading to higher recognition rates under noisy conditions; however, this only works when the noise in the test condition is (close to) equivalent to the contamination noise, and is, therefore, a fairly superficial solution.
Model compensation methods are similar, yet more dynamic. They focus on estimating the noise's energy, after which it is compensated for through a process of parameter modification. The noise estimation step can be solved by a number of algorithms, which won't be further discussed as they are quite technical. However, see Fu-Hua-Liu et al. (1993) for a brief review.

5 Harmonic frequencies as a new set of features.

What follows is an experiment for evaluating the noise robustness of a new set of features called harmonic frequencies (5.1), during which a demonstration of their utility in speech segregation techniques is given.
The general idea is to extract the harmonic frequencies from the TIMIT database (5.2), and to use these to synthesize a new signal (5.3). This process has two variants, called F0 and SC, of which both are applied to the TIMIT database. This thus results in three databases: (i) the original TIMIT database, (ii) its F0 re-synthesized version, and (iii) its SC re-synthesized version; of which all three are independently used by the HTK toolkit to train three sets of HMMs (5.4). These are then evaluated under various noise conditions (5.5).

5.1 Harmonic frequencies

Harmonic frequencies are produced during voiced speech by the rapid opening and closing of the vocal cords. This is produced by the Bernoulli effect, which, in this case, states that the continuously streaming air in the larynx leaves an area of decreased air pressure in its wake. The decrease in air pressure causes the vocal cords to close, which are almost immediately opened again by the outgoing air.
The speed at which the vocal cords open and close determines the frequency of the lowest harmonic; the fundamental frequency. A series of higher frequency harmonics are also produced, which have the convenient property of being multiples of the fundamental frequency.

5.2 TIMIT

The TIMIT (Texas Instruments and Massachusetts Institute of Technology) database was used for training the HMMs. TIMIT is a continuous speech corpus of phonetically-balanced English speech. It contains a total of 6300 sentences, recorded from 630 speakers from 8 major dialect regions of the United States. 1260 of these sentences (the sa1 and the sa2 sentences) promote dialect robustness and are excluded. The remaining utterances are divided into a train (3969 sentences) and a test (1344 sentences) set. Furthermore, the transcriptions are realized by a group of 61 phonemes, which, as proposed by Lee & Hon (1989), are collapsed into a group of 39.

5.3 Auditroy grouping

This subsection provides a high level description of a data-driven technique, similar to those described by Cook (2004), for speech re-synthesis (Vooren, 2011. Unpublished). Ultimately, the method enables speech segregation under multiple speaker conditions whilst simultaneously reconstructing the signal into a noise robust representation.
The general idea is to differentiate the signal into individual frequencies, after which they are re-assembled into individual streams by exploiting the following acoustic cues: proximity in frequency and time, temporal continuity, harmonicity, amplitude and frequency modulation, and on and offset.
The first step in this process is to extract the time-frequency track, which is a simplified version of the spectrum where only the dominant frequencies are represented, which is to say; those that are temporally connected. The features mentioned above are hereby extracted and used with a set of heuristics for the construction of a conflict matrix.
The conflict matrix states the probability that some group of frequencies were induced by the same source; however, determining these probabilities cannot be solved mathematically as there are an infinite amount of solutions, and must therefore be approximated during an iterative, re-estimation process.
As mentioned earlier, two re-synthesis procedures were applied to the data. The difference in the resulting signal can be found at the higher, less energetic frequencies, which are absent in SC and present in F0. The SC method corresponds to the process described above, which thus discards the higher level harmonics during the construction of the time-frequency track, mainly due to their low energy. The F0 method uses the fundamental frequency to reconstruct the entire signal by exploiting the fact that the higher level harmonics are all multiples of the fundamental frequency. Thus, the signal components that

were not extracted in the previous method are now added.
A final remark on the re-synthesis process includes the fact that not all speech utterances produce harmonics. For these unvoiced speech segments, no re-synthesized signal can be constructed. This resulted in a number of silences, which were filled with speech components from the original signal.

5.4 Training and Testing

The three databases were used independently for the training and testing of three sets of HMMs. These were built with HTK (Hidden markov model ToolKit); a free source and highly adaptable toolkit for constructing and manipulating HMMs.
HTK has built in models for feature extraction, training, and testing. Its feature extraction model (HCopy) was used to represent the databases as collections of MFCCs. In total, 13 mel-based cepstral coefficients with e-normalized energy were extracted. Delta and delta-delta coefficients were also added for a total of 39 features.
These features were then used to train three sets (one for each database) of 40 single mixture monophone HMMs with diagonal covariance. This was achieved with HTK's built in modules, which followed, in our case, a three staged process: (i) the initialisation of HMM parameters by HInit, (ii) iterative parameter tuning with the Baum-Welch re-estimation algorithm by HRest, and (iii) further parameter tuning using three loops of embedded Baum-Welch re-estimation by HERest. These models were then tested with HVite, which selects the model that most accurately generates the utterance in question. A transcription file is then constructed and compared with the original transcription so as to evaluate the performance.
A number of test conditions were then constructed by adding pink noise with signal to noise (S/N) ratios of 30, 15, 12, 9, 6, 4, 0, -3, -6, and -9 dB.

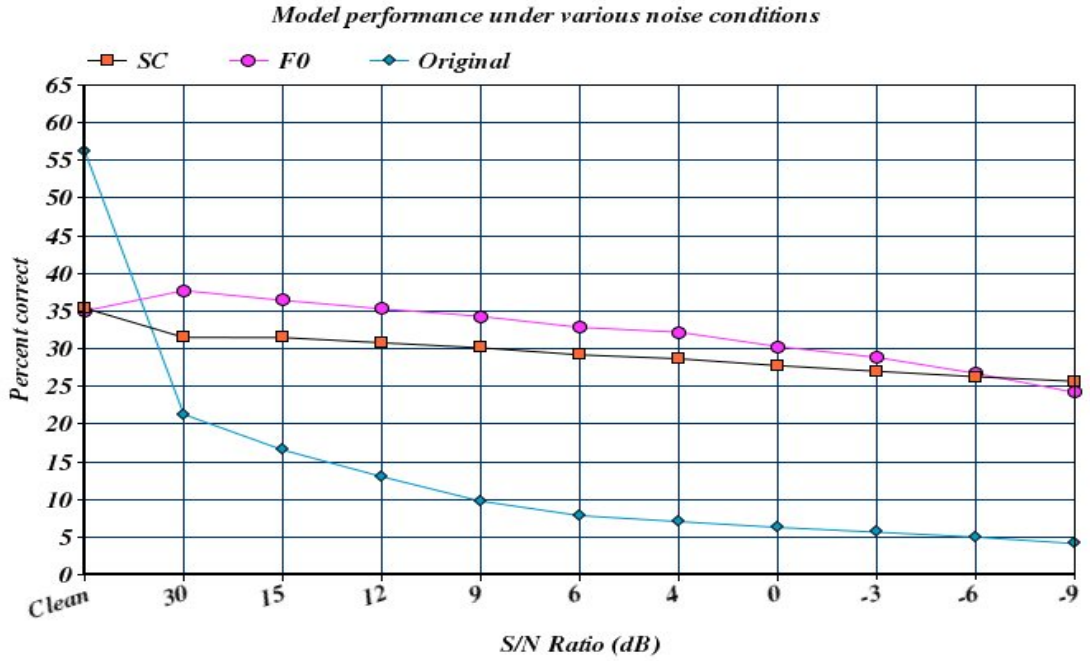
5.5 Results

As can be seen in the graph below, the performance of the SC and F0 models show an almost linear declination under increasing noise conditions, whilst, for a S/N ratio of 30 dB, the original models plummet from 56.2% to 21.28%, all the way down to 4.13% for the -9 condition. SC is fairly stable with a 35.38% recognition rate in the clean condition and 24.2% recognition rate in the -9 condition. F0, although outperforming SC in all noise conditions except -9, is slightly less robust as the performance shows a stronger declination. Finally, both SC and F0 were outperformed by the original models in the absence of noise.
To test whether the results concerning the noise robustness of the re-synthesized models are significant, a confidence interval of 99% was calculated (table 1) where the population consisted of the ten noise conditions. The results indicate that the performance of the re-synthesized models is indeed significantly more robust to noise than the original models. The difference in performance concerning F0 and SC is not significant.

Table 1: Confidence interval

	Original	SC	F0
Lower bound	5.08%	27.12%	28.34%
Upper bound	14.23%	30.57%	35.4%

Figure 1: Model performance under various noise conditions



6 Discussion

This article started with a brief review of our current knowledge of the high-level, functional organisation in the human auditory system. This was an attempt to define a hierarchy of essential components for the construction of an optimal ASR system. Although the human auditory system merely acts as evidence that such systems are indeed realizable, and can therefore not be used to exclude biologically implausible solutions, it is important as it can inspire artificial solutions. Assuming that our current neurological knowledge is, at this conceptual level (excluding implementational details), more or less complete, then the components that were identified ought to be sufficient for the construct of an ASR system with human level recognition rates.

ASR research has thus been found to be closely coupled with neurological findings, albeit at a conceptual level. However, it is often rather specific, whereas the complexity of the auditory system suggests that one should not hope to fully resolve ASR challenges with isolated, domain-specific approaches. Instead, one might expect solutions to be found in complex systems that incorporate and integrate the full assortment of necessary functions.

This is a highly ambitious task, yet some of todays best ASR systems, such as Dragon and

CMUSphinx, have succeded in integrating a wide range of functionalities, with near human-level recognition rates as a result. CMUSphinx-4, for example, deploys a HMM-core with left bigrams, a probablistic language model, noise compensation methods, and noise robust features (Walker et al., 2004), resulting in recognition rates of up to 98%. However, it should be noted that these systems are usualy speaker dependent and are still susceptible to noisy and other sub-optimal conditions.

It should furthermore be noted that, allthough solutions to ASR challenges most likely require the integration of a wide range of functionalities, this does not negate the importance of domain-specific techniques. Instead, it (i) vindicates their low recognition rates, and (ii) stresses the importance of pre-processing techniques, such as the auditory grouping method presented in section 5.3.

With respect to the results presented in section 5.5, it is therefore not unexpected that the recognition rates are so low. However, interesting observations do include the fact that the re-synthesized models were greatly outperformed by the original models in the clean condition, which suggests that a significant amount of essential information resides in the non-harmonic phonemes.

This observation is perhaps partly explained by the fact that, as mentioned earlier, harmonic frequencies are produced exclusively by voiced utterances. One could would therefore expect a large divergence in performance amoungst the voiced and the unvoiced models.

A second interesting observation is the fact that F0 outperformed SC in all noisy conditions except at -9. Allthough the difference was found to be insignificant, visual inspection clearly indicates that F0 is less robust to noise than SC. This can be solely attributed to F0's higher frequency harmonics which are absent in SC.

That these higher level frequencies reduce noise robustness can perhaps be attributed to their low energy levels, since these are more susceptible to noise distortions. This is perhaps why harmonic frequencies are more robust to noise; due to their relatively high energy levels, which enable them to stand out in noisy environments.

To conclude, sound segregation techniques as pre-processing components are biologically plausible. The technique applied here has shown how harmonic frequencies can be exploited so as to both segregate the acoustic environment into acoustic streams, and to transform the input into a noise robust representation. Allthough the time complexity of speech segregation techniques is still considerable this should not deter their development, as these "hardware problems" will most likely be resolved by future advances in other diciplines.

Harmonic frequencies have furthermore been found to be significantly more robust to noise than standard MFCC features. The significance of the results warrents a future research effort for integrating these promising new features with future practical and scientific applications.

References

[1] Bahlmanna, J., Schubotza, R.I.,& Friederici, A.D. (2008). Hierarchical artificial grammar processing engages Broca's area. In *Neuroimage*, Vol. 42(2), pp. 525-534.

[2] Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation for probablistic functions of Markov Processes. In *Inequalities*, Vol. 3,

pp. 1-8.

[3] Békésy, G.V. (1932). ber den Einflu der durch den Kopf und den Gehrgang bewirkten Schallfeldverzerrungen auf die Hrschwelle. In *Annalen der Physik*, vol. 14, pp. 115-125.

[4] Békésy, G.V. (1941). ber die Messung der Schwingungsamplitude der Gehrknchelchen mittels einer kapazitiven Sonde. In *Akust Z.*, Vol. 6, pp. 1-16.

[5] Bellegarda, J.R. (1997). Statistical techniques for robust asr: review and perspectives. *In Proceedings of EuroSpeech*, Vol. 77, pp. 33-36.

[6] Bregman, A.S. (1990). Auditory scene analysis: the perceptual organization of sound. The MIT Press, Cambridge, MA

[7] Butler, R.A., & Belendiuk. (1977). Spectral cues utilized in the location of sound in the median saggital plane. In *The Journal of the Acoustical Society of America*, vol. 61(5), pp. 1264-1269.

[8] Calford, M.B., & Aitkin, L.M. (1983). Ascending projections to the medial geniculate body of the cat: evidence for multiple, parallel auditory pathways through thalamus. In *The Journal of Neuroscience*, Vol 3(11), pp. 2365-2380.

[9] Calvert, A.G., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W.R., Iversen, S.D., & David, A.S. (1997). Activation of auditroy cortex during silent lipreading. In *Science Magazine*, Vol. 276(5312), pp. 593-596.

[10] Cheveign, A.D. (2001). The auditory system as a "separation machine". In *Physiological and Psychophysical Bases of Auditory Function*, pp. 453-460.

[11] Chevillet, M., Riesenhuber, M. & Rauschecker, J.P. (2011). Functional correlates of the anterolateral processing hierarchy in human auditory cortex. In *The Journal of Neuroscience*, Vol. 31(24), pp. 9345-9352.

[12] Cook, P.R. (2004). Real Sound Synthesis for Interactive Applications, A K Peters/CRC Press, Ltd.

[13] Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28(4), pp. 357-366.

[14] Flannery, R. & Butler R.A. (1981). Spectral cues provided by the pinna for monaural localization in the horizontal plane. In *Attention, Perception & Pyschophysics*, vol. 29(5), pp. 438-444.

[15] Yifan, G. (1995). Speech recognition in noisy environments: a survey. In *Speech Communications*, Vol. 16(3), pp. 261-291.

[16] Griffiths, T.D., Bates, D., Rees, A., Witton, C., Gholkar, A., & Green, G.G.R. (1997). Sound movement detection deficit due to a brainstem lesion. In *The Journal of Neurology, Neurosurgery & Psychiatry*, Vol 62(5), pp. 522526.

[17] Guinan, J.J.Jr., Norris, B.E., & Guinan, S.S. (1972). Single auditory units in the superior olivary complex. II: Locations of unit categories and tonotopic organization. In *The International Journal of Neuroscience*, Vol 4(4), pp. 147166.

[18] Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis for speech recognition. In *The Journal of the Acoustical*

Society of America, Vol. 87(4), pp. 1738-1752.

[19] Hirsch, I.J. (1950). Relation between localization and intelligibility. In *The Journal of the Acoustical Society of America*, Vol. 22(1), pp. 196-200.

[20] Ito, M., van Adel, B., & Kelly, J.B. (1996). Sound localization after transection of the commissure of Probst in the albino rat. In *The Journal of Neurophysiology*, Vol. 76, pp. 3493-3502.

[21] Jeffress, L.A. (1948). A place theory of sound localization. In *The Journal of Comparative and physiological psychology*, Vol. 41(1), pp. 35-39.

[22] Johnstone, B.M., Patuzzi, R., & Yates, G.K. (1986). Basilar membrane measurements and the travelling wave. In *Hearing Research*, Vol. 22, pp. 147-153.

[23] Juang, B.H., & Rabiner, L.R. (2004). Automatic speech recognition - A brief historf of the technology development. Rutgers University and the University of California, Santa Barbara.

[24] Kulesza, R.J., Vinuela, A., Saldana, E., & Berrebi, A.S. (2002). Unbiased stereological estimates of neuron number in subcortical auditory nuclei of the rat. In *Hearing Research*, Vol. 168, pp. 12-24.

[25] LeDoux, J.E., Farb, C., & Ruggiero, D.A. (1990). Topographic organization of neurons in the acoustic thalamus that project to the amygdala. In *The Journal of Neuroscience*, Vol 10(4), pp. 1043-1054.

[26] Lee, K., & Hon, H. (1989). Speaker-independent phone recognition using hidden Markov models. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37(11), pp. 1642-1648.

[27] Fu-Hua, L., Stern, R.M., Xuedong, H., & Acero, A. (1993). Efficient cepstral normalization for robust speech recognition. In *Proceedings of DARPA Speech and Natural Language Workshop*, pp. 69-74.

[28] May, B.J. (2000). Role of the dorsal cochlear nucleus in the sound localization behavior of cats. In *Hearing Research*, Vol. 148, pp. 7487.

[29] Mei, L.T. (2009). Cellular mechanisms of auditory processing in the inferior colliculus, an in vivo patch clampstudy. EUR. Prom./coprom.: Feenstra, L. & Borst, J.G.G.

[30] Mohamed, A.R., Dahl, G., & Hinton, G. (2009). Deep Belief Networks for phone recognition. In *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*.

[31] Oertel, D. (1999). The role of timing in the brain stem auditory nuclei of vertebrates. In *Annual Review of Physiology*, Vol. 61(1), pp. 497-519.

[32] Oertel, D., & Young, E.D. (2004). Whats a cerebellar circuit doing in the auditory system? In *Trends in Neurosciences*,Vol. 27(2), pp. 104-110.

[33] Petkov, C.I., Xiaojian, K., Alho, K., Bertrand, O., Yund, E.W., & Woods, D.L. (2004). Attentional modulation of human auditory cortex. In *Nature Neuroscience*, Vol. 7(6), pp. 658663.

[34] Qifeng, Z., Iseli, M., Xiaodong, C., & Alwan, A. (2001). Noise robust feature extraction for asr using aurora 2 database. In *Proceedings of InterSpeech*, Scandinavia, pp. 185-188.

[35] Saint Marie, R.L., Schneiderman, A., & Stanforth, D.A. (1997). Glycine-immunoreactive projections of the cat lateral superior olive: possible role in midbrain ear dominance. In *The Journal of Comparative Neurology*, Vol. 389, pp. 264-276.

[36] Shore, S. (2005). Multisensory integration in the dorsal cochlear nucleus: Responses to acoustic and trigeminal stimulation. In *The Journal of Neuroscience*, Vol 21, pp. 33343348.

[37] Spirou, G.A., Davis, K.A., Nelken, I., & Young, E.D. (1999). Spectral integration by type II interneurons in dorsal cochlear nucleus. In *The Journal of Neurophysiology*, Vol. 82(2), pp. 648-663.

[38] Spitzer, M.W., & Semple, M.N. (1993). Responses of inferior colliculus neurons to time-varying interaural phase disparity: effects of shifting the locus of virtual motion. In *The Journal of Neurophysiology*, Vol. 69, pp. 1245-1263.

[39] Stebbins, W.C. (1980). The evolution of hearing in the mammals. In Popper, Fay, *Comparative studies of hearing in vertebrates*, pp. 421-436, (Springer, New York).

[40] Stevens, S.S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. In *American Journal of Psychology*, Vol. 53(3), pp. 329-353.

[41] Strutt, J.W. (1907). On our perception of sound direction. In *The Philosopher's Magazine*, Vol. 13, pp. 214-232.

[42] Suga, N. & Ma, X. (2003). Multiparametric corticofugal modulation and plasticity in the auditory system. In *Nature Reviews, Neuroscience*, Vol. 4(10), pp. 783794.

[43] Takanishi, A., Masukawa, S., Mori, Y., & Ogawa, T. (1993). Study on anthropomorphic auditory robot continuous localization of a sound source in horizontal plane. In *11th Annual Conference of the Robotics Society of Japan*, Tokyo, Japan, pp. 793-796.

[44] Theodoridis, S., & Koutroumbas, K. (2003). Pattern Recognition: Second Edition, Elsevier, Academic Press.

[45] Tomoko, S., Toru, N., Makoto, K., Ryuichi, K., Ikuro, M., & Zenta, I. (2006). Spectral cues for robust sound localization with pinnae. In *International Confernce on intelligent Robots and Systems*, Beijing, China, pp. 386-391.

[46] Vooren, H. van de. (2011). Titel unknown. Groningen University, Groningen, the Netherlands.

[47] Wang, D. & Brown, Guy J. (2006). Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. IEEE Press/Wiley-Interscience.

[48] Warr, W.B. (1982). Parallel ascending pathways from the cochlear nucleus: Neuroanatomical evidence of functional specialization. In *Contributions to Sensory Physiology*, Vol. 7, pp. 1-38.

[49] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., & Woelfel, J. (2004). Sphinx-4: A Flexible Open Source Framework for Speech Recognition. *Sun Microsystems Technical Report*, No. TR-2004-139.

[50] Wepsic, J.G. (1966). Multimodal sensory activation of cells in the magnocellular medial geniculate nucleus. In *Experimental Neurology*, Vol. 15(3), pp. 299-318.

[51] Winter, I.M., & Palmer, A.R. (1995). Level dependence of cochlear nucleus onset unit

responses and facilitation by second tones or broadband noise. In *The Journal of Neurophysiology*, Vol. 73(1), pp. 141159.