Perceptual Fusion of Noise and Complex Tone by Means of Amplitude Modulation

Pär Johansson

Master's Thesis
Department of Speech, Music and Hearing
The Royal Institute of Technology, Stockholm

Abstract

This report investigates pulse-synchronised amplitude modulation as a method for perceptual fusion of noise and tone. It is shown that when the noise is amplitude modulated by the waveform envelope of the tone, higher amplitude modulation depth yields stronger fusion. In addition, the evidence suggests that fusion remains constant for tone frequencies from 150 to at least 600 Hz.

Sammanfattning

Det är numera välkänt att olika former av brus är en väsentlig del av klangfärgen hos de flesta akustiska instrument, vilket ofta förbises vid konstruktionen av elektroniska och digitala musikinstrument. Forskning inom detta område har visat att bruset tillför en ökad realism till syntetiserade instrument, men också att det inte är tillräckligt att bara mixa en ton med vitt brus för att vi skall uppfatta de båda komponenterna som en sammansatt klang – en lyckad sammansmältning eller perceptuell fusion kräver att bruset och tonen är korrelerade. Ett sätt att åstadkomma detta är att amplitudmodulera bruset med tonens frekvens. Detta är ett fenomen som uppstår naturligt i bl a röst, blås- och stråkinstrument. Denna uppsats behandlar dels hur modulationsdjupet påverkar fusionen, dels hur fusionen varierar med tonens frekvens.

En enkel modell av ett instrument med en bruskälla skapades, där tonens frekvens och ljudnivå, brusets ljudnivå och modulationsdjupet kunde variera. För att pröva hypoteserna 1) fusionen ökar med modulationsdjupet och 2) fusionen minskar med frekvensen, gjordes lyssnartest vid frekvenserna 150, 300 och 600 Hz. Deltagarnas uppgift var att rangordna fem ljud producerade av modellen ovan. De fem ljuden hade modulationsdjupen 0, 25, 50, 75 och 100%.

I genomsnitt rankades 0% modulationsdjup lägst vid alla frekvenser, 50% modulationsdjup högst vid 150 Hz och 75% modulationsdjup högst vid 300 och 600 Hz. En analys av rangordningarna gjordes med Friedmans test som visade signifikanta* skillnader vid 300 Hz och signifikanta*** skillnader vid 600 Hz, vilket innebär att försökspersonerna med stor sannolikhet kunde särskilja mellan olika modulationsdjup. Detta stöder den första men inte den andra hypotesen, eftersom det inte verkar vara svårare att höra skillnader vid högre frekvenser, vilket troligen gäller även när hänsyn tagits till den inlärning som eventuellt förekom.

Tidigare studier har visat att fusion är ett resultat av synkroniseringen mellan tonens och brusets vågformer, vilket skulle betyda att fusionen minskar med ökande frekvens då synkroniseringen blir svårare att uppfatta. Då denna undersökning inte stöder att fusionen minskar mellan 150 och 600 Hz, finns det anledning att ifrågasätta detta antagande. Fler studier är nödvändiga för att klarlägga mekanismerna bakom fusion av ton och brus.

Contents

1. Introduction

It is now well understood that noise is an integral part of the sound of almost every acoustic instrument, especially woodwinds, brass, strings and voice. Noise can also be considered a source for amplitude and frequency modulations, e. g. the shimmer, jitter, flutter and wow of the human voice. This is often overlooked in conventional synthesisers, where a simple sample-and-hold waveform may be the only random source available. An exception is the analogue modular synthesiser, with its inherent variation due to noise and its many modulation possibilities, and indeed, digitally emulating a modular synthesiser is as difficult as synthesising an acoustic instrument, but this is due not only to the deliberate design of the synthesiser, but also to the inherent instability of analogue electronics.

Simply adding noise to a synthesised instrument does not make one believe that one is listening to an acoustic instrument. Rather, the sound tends to split into separate noise and tone streams. To prevent this, the tone and noise have to be correlated in some way. In this thesis, I study pulse-synchronised amplitude modulation as one method for correlating noise and tone. This is a natural phenomenon that occurs in many instruments, including voice, and is believed to 'fuse' the noise and tone perceptually. I also examine if it is possible to perceive different degrees of fusion, and if fusion varies with frequency.

The thesis is organised as follows: I begin by reviewing some previous work and discuss some of the possible conclusions of that research. I then describe the simple synthesis model I developed for use in the listening tests. A discussion of the listening tests and an analysis of the test data follows. I conclude with a short summary and some directions for further research.

2. Previous work

2.1. Perceptual fusion

Perceptual or spectral fusion[1] belongs to a large area of research, known as *auditory scene analysis*, which deals with different types of integration and segregation of sounds. While several researchers have conducted studies on perceptual fusion and related topics, few of these researchers were concerned with the fusion of noise and tone. The studies most relevant to this thesis are reviewed below. For more extensive reviews, see Bregman (1990) and Moore (1997).

The important auditory cues for perceptual fusion can be found in McAdams (1984), Bregman (1990) and Moore (1997). These include correlated amplitude and frequency variations, harmonicity, spectral shape and spatial location. There are several types of correlated amplitude variations, the most important ones being onset-offset synchrony—sometimes regarded as a separate correlation type—and periodic or random amplitude modulation. However, these cues do not tell us what perceptual fusion really is, nor how to distinguish it from related phenomena such as masking or layering—defining fusion is far from trivial. While a strict definition of fusion is not a prerequisite for this thesis, we can not do without a theoretical discussion.

Fales and McAdams (1994) provides the following definition:

> [… W]e might say that noise fusion exists if two conditions are met: (1) the overall timbre of the sound must demonstrate the mutual influence of the two components (so that […] we have a noisy flute […] as opposed to a flute […] in the presence of noise), and (2) when the sound is analytically divided into its two components, both the noise alone and the tone stripped of the noise must sound discernibly different in the absence of the other.

This intuitive definition is somewhat vague, and I would like to point out some of its problems. Condition (1) is not sufficient to explain the difference between a noisy flute and a flute in the presence of noise. Indeed, if it were sufficient, it would also be a sufficient definition of fusion! This gives the definition a circular character.

[1] I prefer the more general term 'perceptual fusion' (or simply 'fusion') to 'spectral fusion', because 'spectral' gives the impression that spectral cues like harmonicity and spectral shape are more important than temporal or spatial correlation.

Conditions (1) and (2) both hold for sounds composed of unrelated components partially masking each other. It follows that (1) and (2) together does not provide a sound definition of fusion. In addition, how to 'analytically divide' the sound in its components is not a trivial task. In fact, these components are often produced by the same source, and thus the 'mutual influence of the two components' is simply the result of this common origin, and not a somewhat magical interaction. This point will be further elaborated on below.

The Fales-McAdams definition tries to capture the notion that the fused sound is an *emergent* quality of the combined sources, in contrast to *augmented* sounds, where the timbre of one of the sources changes in the presence of the other, which remains almost unchanged.[2] In contrast, *layered* sounds do not influence each other in this way. Most authors seem to agree that timbre change is a necessary result of fusion, and some have used perceived timbre change as a measure of fusion (e. g. Hermes, 1991, see below). Interestingly, Fales and McAdams did not use this approach, but asked the subjects if they could hear the tone and the noise separately.

However, even in the case of layered sounds, like the musical bow (Fales and McAdams, 1994, see below), there is at least spatial correlation and onset-offset synchrony between the components,[3] which is not the case for a sound in the presence of background noise. Therefore, we may conclude that layered sounds do sometimes fuse in a way that segregates them from other sounds. On the other hand, in even the strongest fused sounds, complex harmonic tones, human perception can single out individual harmonics, as already Helmholtz noted. Hence, the perception of fusion does not seem to be categorical. Rather, the characterisations 'layered', 'augmented' and 'fused' can be viewed as different levels on a relative and more or less continuous scale. Timbre change of individual components thus becomes a feature of the higher levels of this scale, but should not be used to define fusion. This obviously means that tests based on timbre change exclude the lower levels of fusion, and thus can not provide definite evidence for fusion except at high levels.

I do not argue that timbre change is irrelevant for fusion. Rather, a change in fusion *is* a timbre change, or, put differently, fusion is one of the many aspects of timbre. A test based on timbre change may fail to accurately identify fusion related timbre changes from timbre changes that are artefacts of the test itself. For a discussion of this with respect to masking, see Hermes (1991). In order to avoid a definition involving timbre, a concept hardly less elusive than fusion, I suggest that in a highly fused sound, the individual components must be clearly audible, that is, one is able to hear them individually if one directs one's attention to them. This means that neither of them may mask the other completely, but they may mask each other if they still retain their individuality. The sources must also be correlated in a manner listeners can perceive (not necessarily consciously). The degree of correlation is context-dependent. In particular, successful noise-tone fusion probably demands a very strong correlation, and one method for achieving this is pulse modulation of the noise at the tone frequency.

2.2. Noise in acoustic instruments

Chafe (1990) investigated the addition of pulsed noise to physical models of reed and bowed string instruments. This research resulted in two patents, Chafe (1992) and Chafe (1996), that may have been used for commercial synthesisers. The pulsed noise in woodwind instruments is a result of the airflow turbulence at the reed aperture being pulse-modulated as the reed(s) beat(s). Rocchesso and Turra (1993) reported a "dramatic increase" of realism when noise was added to their physical model of the clarinet reed. Chafe's research (1993) also show that while flute noise is more continuous than reed or string noise, noise pulses are present and that they are spectrally weighted toward higher frequencies.

In bowed string instruments, irregularities in the slip-stick motion of the bow lead to the production of noise. McIntyre et al. (1981) showed that this noise is a result of *differential slipping*, when some but not all of the bow hairs release during the sticking phase. The noise becomes pulsed due to the abrupt

[2] This distinction is based on Gregory Sandell's research on instrument blending in orchestration. For references, see Sandell (2000).

[3] These factors are clearly important in a musical context, but in a listening test, where the subject's focus of attention is directed at other stimuli, they may play a relatively minor role.

slipping of the bow (Chafe, 1990). Schumacher (1993) claims that the bow noise is also related to the "ghostly" subharmonics phenomenon discussed by McIntyre et al. (1981).

In wind instruments, the turbulence noise is filtered by the resonances of the pipe. Since these resonances are inharmonic, the peaks of the noise spectrum are located between the harmonics of the tone in instruments with coupling between excitation and pipe. This becomes apparent especially for high harmonics. See Sundberg (1966) and Hirschberg and Verge (1995) for examples.

Fales and McAdams (1994) studied the fusion and layering of noise in three African instruments: the mbira (sansa), the bamboo flute and the musical bow. Listening tests were also made on three parameters related to fusion, namely, relative intensity, noise centre frequency and bandwidth. The stimuli consisted of single sine tones mixed with bandpass filtered noise. Listeners were asked to select the one of the following statements that matched their perception: "Tone heard separately: I am sure (1); I am fairly sure (2); I am not sure (3); Tone NOT heard separately: I am not sure (4); I am fairly sure (5); I am sure (6)". According to Fales and McAdams, levels 3 and 4 indicate fusion and levels 5 and 6 indicate that the tone may be masked by the noise.

The listening tests showed that fusion occurred in the low register (400 Hz) when the intensity of the noise was between -7 and -1 dB relative to the intensity of the tone. For the high register (1000 Hz), the corresponding levels were between -3 and +4 dB. When the noise bandwidth was increased from 50 to 75 Hz, these intensity thresholds decreased slightly for both low and high registers, and then remained constant for larger bandwidths. In the centre frequency test, both intensity and bandwidth were held constant such that no masking could occur. The offset between tone frequency and noise centre frequency was varied. Listeners' responses showed that fusion decreased with increasing offset.

As Fales and McAdams remark, it is difficult to generalise these results to the fusion of a complex tone and noise. Due to masking phenomena, the relative intensity of the noise source seems to be the most important parameter. However, the noise and tone spectra must overlap, or the noise is perceived as a layered sound, as was the case of the high frequency noise of the musical bow (see below).

In the bamboo flute, we see the same doubled spectral peaks (the harmonics and in-between noise peaks) as discussed above. The strongest noise peaks coincide with the first and third harmonics. The fusion between noise and tone is weaker than that in the mbira. Fales and McAdams do not discuss the possible influence of pulsed amplitude modulation on the noise-tone fusion.

The mbira consists of metal or wooden lamellae attached to a resonating box. These lamellae are plucked with the thumbs or forefingers of both hands. The instrument often has metal strips wrapped around the lamellae or bottle caps or other objects attached to the resonating box, which gives a rattling, noisy timbre. There may also be some noise present even without those devices.

In the studied mbira, noise is produced by bottle caps. The fusion between noise and tone is stronger than the flute, maybe because the partials are strongly inharmonic and do not fuse well together. Although Fales and McAdams do not mention this, I, having listened to some mbira music, believe that the bottle caps vibrate at a rate partly dependent on the frequencies of the plucked lamellae. This may contribute to fusion. As in the flute, the noise is roughly centred around the partials.

The musical bow used for Fales and McAdams's experiment is a struck string instrument with a single string and a gourd resonator held against the player's chest.[4] The string is far from ideal, and the partials are inharmonic, though less so than in the mbira. Noise is produced by metal 'clackers' attached to the resonator. In the analysis process, all of the frequencies above 6 kHz were separated from the rest of the sound, since no audible partials could be found in this region. This high frequency noise was perceived as a layered sound. The noise left with the partials, being roughly centred around the partials, was similar to the noise of the flute and the mbira.

[4] The musical bow is a widespread instrument that comes in several varieties. The player's mouth is often used as a resonator. For details, see Nketia (1974), still a good introduction to the music of Africa.

2.3. Some possible functions of noise in acoustic instruments

Is there a reason for having a strong noise component in certain instruments? McAdams (1984) suggests that onset noise may mask the differences in onset times for partials of musical instruments, thus increasing their fusion.

In the African instruments studied above, the presence of noise is particularly noticeable, as if they were designed to be noisy. Fales and McAdams suggest that the presence of noise peaks in the bamboo flute spectrum affects the pitch of the instrument. The somewhat inharmonic partials should make the flute produce a lower perceived pitch than if the partials had been perfectly harmonic, but the noise peaks, located above the partials, pull the pitch upwards. Listeners reported that flute sounds with the noise part removed did have a lower pitch. They also reported that the presence of noise reduced the 'roughness' of the tone. In the mbira and the musical bow, noise compensated for inharmonicity in a similar way and contributed to the fusion of the partials.

Being instruments difficult to tune, the mbira and other idiophones may benefit from additional noise when used in ensemble playing. Fales and McAdams reported that when two mbira musicians did not succeed in tuning their instruments together, they attached more bottle caps to the resonating boxes.

2.4. Noise in voice

Adding pulsed noise to synthesised speech provides for increased realism and extends the available speaker styles with, for example, breathy or hoarse voice types. Carlson, Granström and Karlsson (1991) used the glottal flow waveform to modulate the noise in the GLOVE speech synthesiser, which slightly improved the realism of the synthesised female voice.

Hermes (1991) mixed lowpass filtered pulse trains and highpass filtered noise bursts to simulate breathy vowels. The noise burst rate was equal to the pulse train frequency, 125 Hz, and the phase difference between pulses and noise bursts varied between 0 and 2π. To measure the segregation of pulse trains and noise bursts, subjects were asked to adjust the level of a comparison source so that its loudness was equal to that of the noise bursts of the synthetic vowel. To measure the integration, subjects were asked to adjust the high-frequency content of a pulse train so that it matched the timbre of the vowel.

The listening tests showed that the loudness of the noise bursts was lowest and the high-frequency content of the vowel was highest when pulses and noise bursts were synchronised. Hermes' explanation is that when the highpass filtered noise is better integrated into the vowel, it affects the vowel's timbre so that its high-frequency content increases.

In a somewhat similar experiment, Skoglund and Kleijn (1998) showed that the detection of phase differences between noise bursts and glottal pulses is frequency dependent. Phase differences are easier to detect at lower frequencies (100 Hz) than at higher (200 Hz).

The ability to detect rapid amplitude modulation (AM) may be important for understanding speech. Menell, McAnally and Stein (1999) showed that dyslexic listeners are less sensitive to AM than control listeners. In addition, their study shows that the sensitivity to AM is frequency dependent. At 80 Hz, the detection threshold for modulation depth was significantly lower than that for 160 Hz, which, in turn, was significantly lower than that for 320 Hz.

2.5. The pitch of SAM and iterated rippled noise

While the long-term spectrum of AM noise is white and does not depend on the modulation frequency f_m, the short-term spectra do contain information that could be used to detect the modulation frequency in that a momentary peak at f_c is reflected at $f_c - f_m$ and $f_c + f_m$. It has been speculated that this may be the cause for the vague pitch sensations elicited by AM noise. For a review of this research, see Burns and Viemeister, 1981.

However, Burns and Viemeister's studies of sinusoidally amplitude modulated (SAM) white noise show that the pitchlike sensation is mediated by temporal information rather than by short-term spectral information. Pitch was defined as the perception that conveys melodic information, and listening tests were based on listeners' ability to detect melodies played with SAM noise with varying modulation frequency. Listeners could obtain pitch information even when the noise was bandpass filtered between 9 and 11 kHz. For modulation frequencies above 500 Hz, the pitch sensation rapidly vanished.

Iterated rippled noise (IRN) is produced by comb filtering white noise and produces a similar pitch sensation as SAM noise. The results of Yost, Patterson and Sheft (1998) imply that the temporal properties—the fine structure or, less likely, the envelope[5]—of the IRN waveform are responsible for the perceived pitch. Listeners were also able to discriminate between flat-spectrum noise and IRN highpass filtered at 8 kHz, but these discriminations were not dependent of the IRN pitch.

The central auditory system (CAS) has channels dedicated to detecting modulation that may be the basis for periodicity pitch and rhythm perception. The AM detectability curve has maxima at 3 and 300 Hz (McAngus Todd and Lee, 1998), but the sensitivity to AM decreases with modulation frequency and has completely vanished at 1000 Hz (see the sections on phase and temporal modulation transfer functions in Moore, 1997). Thus, these AM channels may be responsible for the perception of pitch elicited by SAM noise, and may be important for noise-tone fusion at lower frequencies. However, it is not obvious that the CAS AM channels can detect IRN pitch. Since there are no periodic envelope fluctuations in IRN, Yost, Patterson and Sheft (1998) suggest that the mechanism used may be based on autocorrelation.

2.6. Analysis methods

The recent interest in noise makes the need for better analysis methods apparent. The *phase vocoder* (Dolson, 1986) is a well-known tool for analysis, resynthesis and high quality time and pitch scale modification of audio signals based on the short-time Fourier transform (STFT). However, the phase vocoder does not represent noise or transients well because of the inherent limitations of the Fourier analysis. In addition, the rather long analysis window (about 30 ms) and the overlap-add resynthesis destroy the time location of transients, making sharp attacks lose their percussiveness. This artefact, known as *transient smearing*, is clearly audible when the phase vocoder is used for larger time-scale modifications. For details and a solution to this problem by synchronising the analysis window with the transient events, see Masri and Bateman (1996).

Phase vocoder based methods use the phases of the individual STFT channels to calculate the exact frequencies. This process destroys the phase alignment across the channels, which leads to artefacts known as *reverberation*, *phasiness* or *loss of presence*. For a discussion of this problem and possible solutions, see Laroche and Dolson (1999). For noisy acoustic instruments, where the phase synchronisation or similarities in waveform fine structure between tone and noise are important, both transient smearing and reverberation may be detrimental.

In *sinusoidal modelling* (McAulay and Quatieri, 1986), sinusoidal frequency trajectories are extracted from the STFT data. Many tools used in electroacoustic music are based on the phase vocoder or sinusoidal modelling. Sinusoidal modelling is more computationally expensive than the phase vocoder and suffer from various transient artefacts and extraneous partials (although not from reverberation), but has the advantage of being able to remove selected time-varying frequency components from the signal, creating a residual. In the ideal case, this residual contains only the noisy and transient parts of the sound, but in reality it also contains the errors of the modelling (Goodwin, 1996).

New methods have been proposed that combines sinusoidal modelling with residual modelling: *deterministic plus stochastic* models. For an in-depth discussion of these, see Serra (1997). A good example of this type of extensions is Serra's own *spectral modelling synthesis* (SMS), where the residual is calculated as the difference between the original sound and the sinusoidally resynthesised signal. This residual is then analysed and resynthesised using time-varying, filtered white noise (Serra and Smith, 1990). In a further development of SMS reported by Verma and Meng (2000), the

[5] The envelope was produced by half-wave rectifying and lowpass filtering the waveform. The authors claim that this type of envelope extraction is the only one compatible with the auditory processing of IRN.

transients are extracted from the residual or from the original signal and analysed with the discrete cosine transform (DCT). Because of the duality of the time and frequency domains, the DCT of a transient in the time domain is actually a slowly varying cosine in the frequency domain. This DCT spectrum is analysed with the STFT and then resynthesised.

A rather different, parametric approach was used by Richard et al. (1993). They analysed the stochastic component with linear predictive coding (LPC), and resynthesised it in the time domain using *random formant waveform synthesis* (RFWF), a descendant of the FOF (*Formant d'Onde Formatique*) systems developed by Rodet and co-workers (Rodet, 1980). The formant parameters for the RFWF synthesis were provided by the LPC analysis, and the time varying amplitude was captured with a Hilbert transform followed by low pass filtering and peak extraction. This method was reported to achieve very good results for all kinds of musical noises.

In SMS and related systems, there is often a lack fusion between the deterministic and stochastic parts, and the deterministic part can sound 'metallic.' These artefacts may be the result of a lack of detail in the sinusoidal trajectories. While the more modern wavelet based methods does solve the transient problem, they suffer from the same problem as STFT based methods when used for stochastic signals. For a good overview and a comparison of different methods, see Masri et al. (1997 a and b).

2.7. Discussion of previous work

The studies reviewed above show that noise-tone fusion is not only dependent on the cues for perceptual fusion in general, but also on relative levels, spectral overlap and phase synchrony. Thus, for AM noise, possible cues for fusion are phase synchrony, spectral and temporal information. As shown by the experiments on SAM noise, short-time spectral information is present, but it is the temporal information that is used for detecting the pitch. That temporal information is more important for noise pitch is confirmed by the experiments on IRN.

If phase synchrony is the most important cue, fusion should decrease with frequency. But it is also possible that the weak AM noise pitch itself contributes to fusion, which means that fusion is possible at somewhat higher frequencies. In that case, the pitch sensation elicited by the resonance of wind instrument pipes could also be an important cue, but whether it is the spectral or temporal properties of the waveform that is used by the auditory system is not clear and must be studied empirically.

The noise in acoustic instruments and voice is not always produced by an independent source. In wind instruments and strings, noise and tone have the same source, a source that exhibits chaotic behaviour to a degree. The noise and tone have similar overall envelopes and perhaps also correlated variations in amplitude at the waveform level that may contribute to fusion. Taking this chaotic element in account in the excitation source of a physical model thus should yield more convincing results than simply adding a separate pulsed noise source to an existing model, and this is indeed shown by the research of Chafe and Rocchesso and Turra. However, for research in fusion and sound processing targeted to electroacoustic music, one would like to control the noise and tone separately. This is not a problem in synthesis from scratch where a simple source-filter model may be used, but for resynthesis purposes, the traditional analysis tools suffer from the problems described above.

In traditional musicology, pitch has been considered as the most important structural element and timbre a result of the mixing of more or less harmonically related tones. This may explain why music acoustics hitherto have not directed much of its attention to noise, and have not questioned the traditional Fourier based methods. Treating noise as a separate source that can be excluded at will is motivated partly by that its importance for pitch and timbre is supposed to be negligible. But, as seen above, noise is an important timbral component, and the evidence presented by Fales and McAdams show that noise may influence the overall pitch of instruments.

The difficulties of traditional methods originate from the fact that deterministic (sines or wavelets) and stochastic (noise) signals are mathematical models that do not capture the intrinsic properties of sound and its production. They are the two extremes of a continuum that ranges from a single frequency to white noise. Composers of electroacoustic music have long been aware of this, and the sound typologies of Schaeffer (1966) and Smalley (1986) reflect this knowledge, but, unfortunately, these typologies are not well suited for scientific purposes.

Ultimately, if Fourier and wavelet based methods do not also take in account the non-deterministic behaviour of sound, they may be a cul-de-sac. Masri (1996) therefore proposes a new analysis model, based on *noisy partials*, i. e. partials that are chaotic to some extent. In his model, based on random frequency modulation, the amount of chaos can be specified in percent for each partial from 0% (sine) to 100% (white noise).

3. A model of a noisy instrument

A model of an instrument based on Hermes' research was developed in Aladdin Interactive DSP 1.3,[6] an interactive, real-time signal processing package. The goal was to produce sounds that sounded natural, while not exactly like any acoustic instrument.[7] As stated above, a physical model with chaotic excitation would be a timbrally superior synthesis engine for noisy instruments. But for research purposes, the advantage of a noise source that can be independently filtered and controlled is considerable. In this study, I have therefore chosen a more traditional source-filter model.

The model consists of a tone and a noise source. The tone is generated by half-wave rectifying and lowpass filtering a sine wave. The pseudo-gaussian white noise is produced by adding three independent, rectangularly distributed white noise sources. The resulting noise is amplitude modulated directly by the tone envelope and then bandpass filtered to form a realistic timbre together with the tone.[8] To adjust the phase difference between the AM noise and tone, a variable delay line is included.[9]

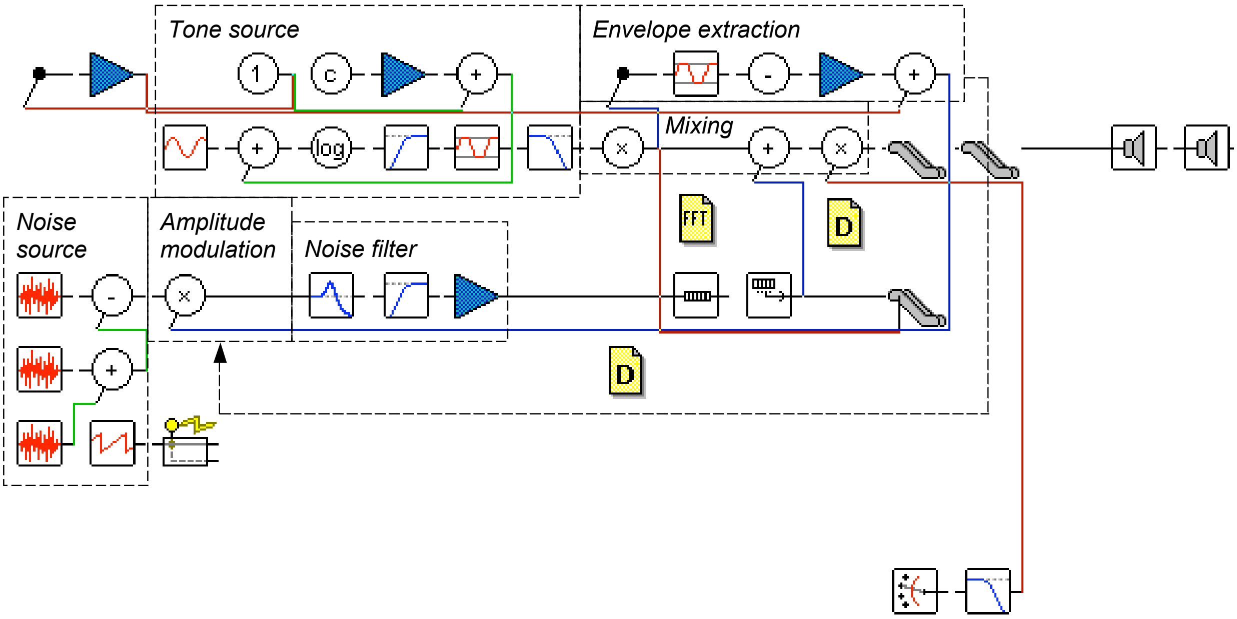


Figure 3.1. The Aladdin patch.

The source file is given in Appendix 1.

[6] Hitech Development AB, `http://www.hitech.se/development/products/aladdin.htm`.

[7] One of the subjects participating in the experiment (see below) remarked that the stimuli sounded like a "bad organ pipe," probably because of the relatively high noise level.

[8] If the filtering were performed before the modulation, the AM noise would no longer be white due to the resulting sidebands.

[9] Phase differences were always set to 0 in the experiment reported below.

4. Experiment

The purpose of the experiment was to determine how the AM depth affects fusion and if fusion varies with frequency. It seems reasonable to believe that increasing AM depth should yield increasing fusion, and if fusion is strongly dependent on phase synchrony or AM detection, one would suspect that fusion should decrease with increasing frequency. Hence, differences in fusion between different AM depths should be harder to detect at higher frequencies.

Most researchers have not asked the subjects directly about their perception of fusion. Instead, they have resorted to more indirect methods. I have chosen the opposite path. If fusion indeed is a natural perceptual category, like pitch or loudness, subjects should be able to compare sounds based on this percept, even if they are initially unfamiliar with the concept of fusion. A disadvantage with this approach is that the data are more difficult to analyse.

4.1. Subjects

12 subjects participated in the experiment. All had some musical experience[10] and none reported having any serious hearing problems. The subjects were not paid for their time.

4.2. Stimuli

The stimuli consisted of three sets of five samples of a complex tone with added bandpass filtered, pulse amplitude modulated gaussian white noise. The three sample sets had average frequencies of 150, 300 and 600 Hz, and the AM depths of the five samples were, respectively, 0, 25, 50, 75 and 100% of the tone level (see fig. 4.1a-c). All samples were created with the Aladdin model described above. The sounds were played on a Soundblaster 16 sound card with 16.0 kHz sampling rate and 16 bit resolution and presented though AKG headphones.

For each subject, each set was randomly selected from a set of eleven samples with different frequencies (140, 142, …, 160 Hz for the first set, 280, 284, …, 320 Hz for the second and 560, 568, …, 640 Hz for the third). The motivation for this was that if all stimuli have the same frequency, subjects might listen too closely to small differences in overall timbre instead of focusing on fusion.

Since informal preliminary listening tests showed that the relative amplitude of the noise is crucial for fusion (see also Fales and McAdams, 1994), perhaps due to masking, noise levels were subjectively adjusted by the author to ensure that all samples had the same perceived tone-to-noise ratio. A different approach would have been to compare the weighted power spectrums of tone and noise, and adjust the noise level for each sample so that the relative amplitudes were equal. To do this in an even more rigorous manner, one would have to take masking effects in account. However, this would still not be enough to ensure that the *perceived* relative levels were the same, since the pulse-synchronisation may result in interactions between noise and tone that would not be compensated for. The less laborious subjective method thus seemed good enough.

[10] Musical experience was not necessary for participating in the experiment. However, I believe that musically trained people have developed somewhat similar conceptions of the terms ‘noise’ and ‘tone’ and thus should be able to intuitively understand the concept of fusion.

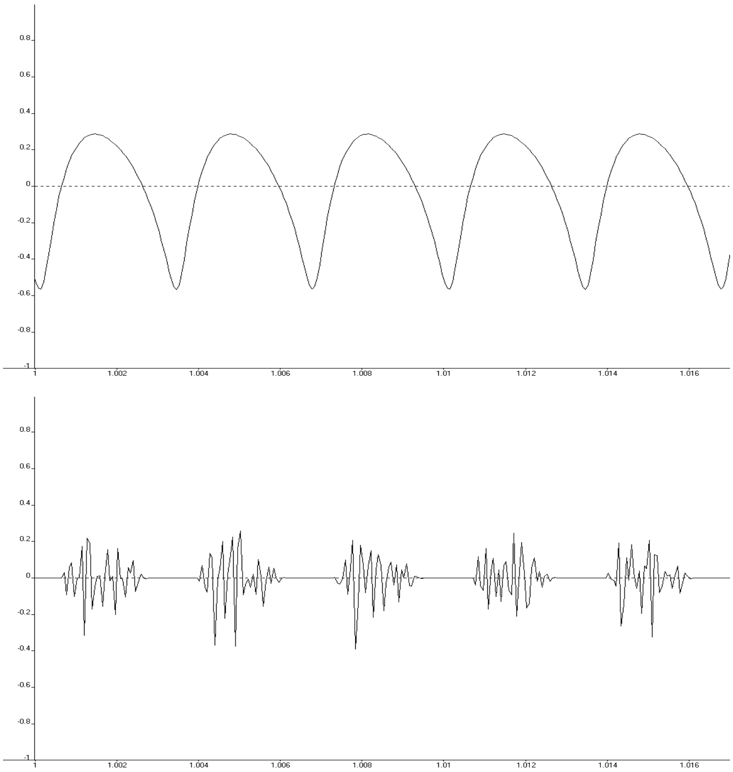


Figure 4.1a. Tone and noise (100% AM depth) waveforms, 300 Hz.

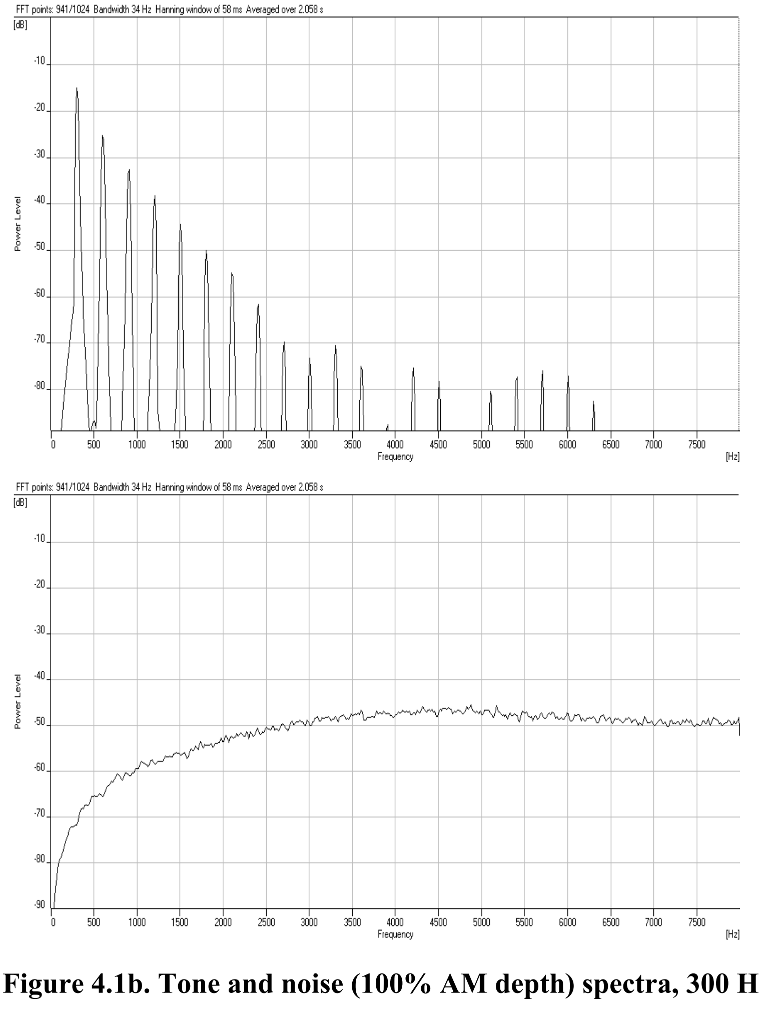


Figure 4.1b. Tone and noise (100% AM depth) spectra, 300 Hz.

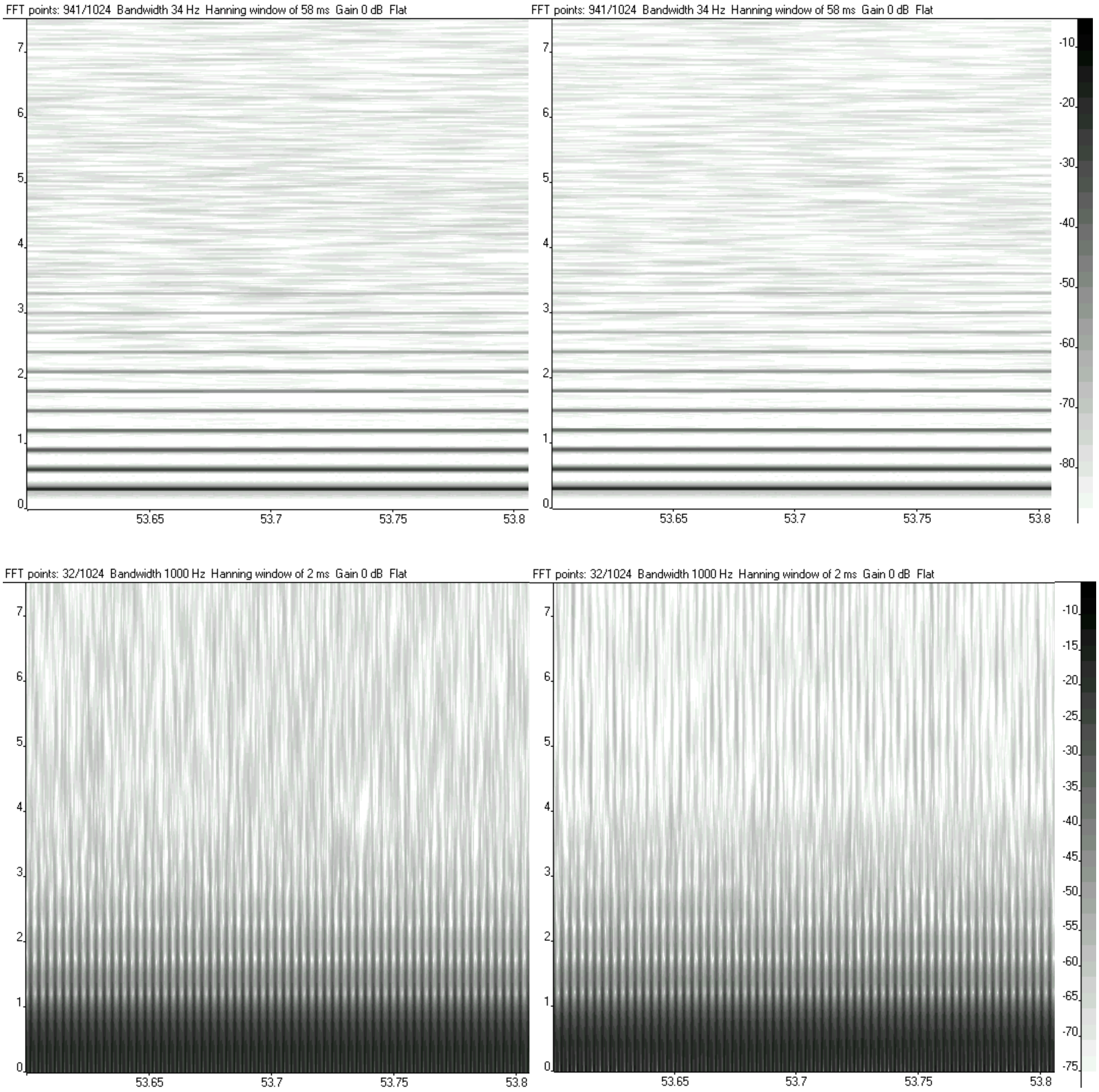
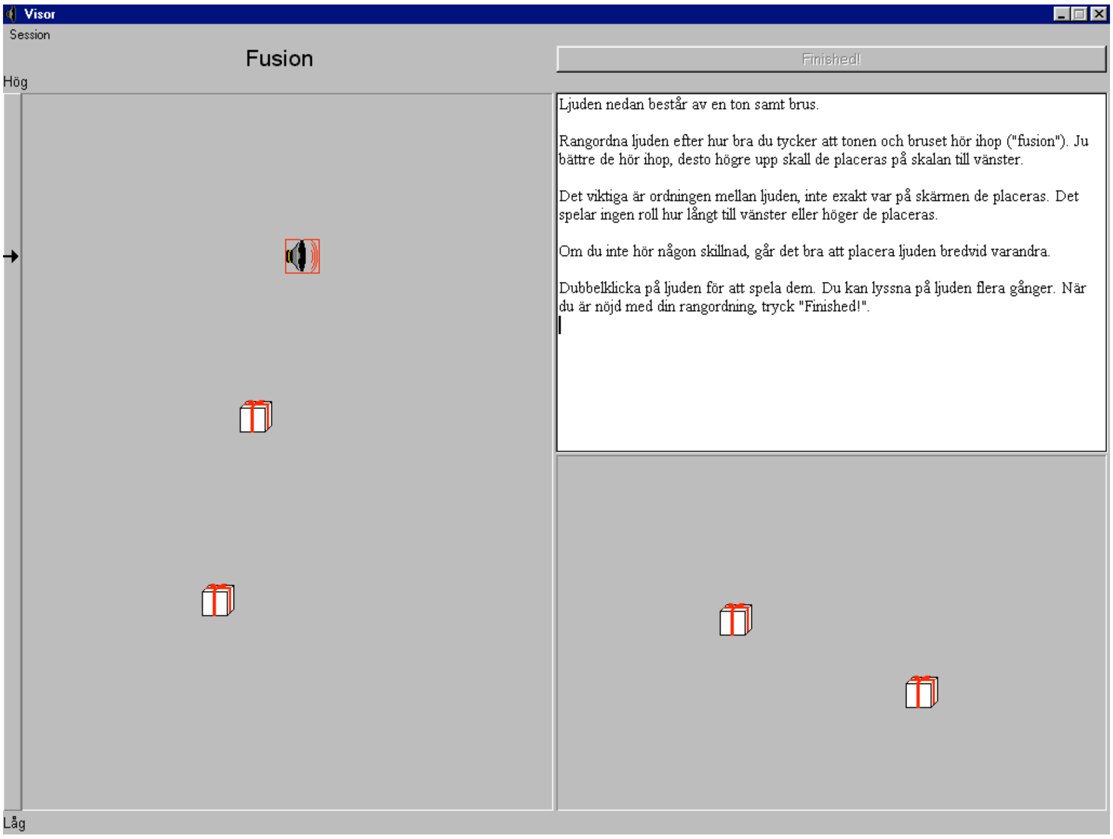


Figure 4.1c. Spectrograms for a 300 Hz sample, 0% (left) and 100% AM depth (right). (Note that there is no difference between the narrow-band spectra for the different depths.)

4.3. Procedure

The subjects were asked to rank the samples in each set according to their perception of the fusion of noise and tone. Ties were allowed. No time limit was set, and the subjects could listen to each sample several times. The ranking was done with Visor,[11] an application where the subjects play and then place the sounds in the preferred order (see fig. 4.2). Test instructions were given in Swedish, since all of the subjects spoke Swedish as their mother tongue. The original text and an English translation are given in Appendix 2.



Ljuden nedan består av en ton samt brus.

Rangordna ljuden efter hur bra du tycker att tonen och bruset hör ihop ("fusion"). Ju bättre de hör ihop, desto högre upp skall de placeras på skalan till vänster.

Det viktiga är ordningen mellan ljuden, inte exakt var på skärmen de placeras. Det spelar ingen roll hur långt till vänster eller höger de placeras.

Om du inte hör någon skillnad, går det bra att placera ljuden bredvid varandra.

Dubbelklicka på ljuden för att spela dem. Du kan lyssna på ljuden flera gånger. När du är nöjd med din rangordning, tryck "Finished!".

Figure 4.2. Visor.

I argued above, that fusion is a continuous rather that than a categorical percept. One may of course question, if this experimental set-up does not favour this unproved assumption by forcing the subjects to rank the sounds. But all experiments influence the subjects' choices in some way—that is, after all, their purpose. By allowing ties I allow for more flexibility, but, as we shall see below, this flexibility was seldom used.

[11] Tolvan Data, `http://www.hitech.se/development/products/spruce/listtest.htm`.

4.4. Results

All data are shown in Appendix 3. The average ranks and standard deviations are shown in figs. 4.3-5. Higher rank indicates stronger fusion.

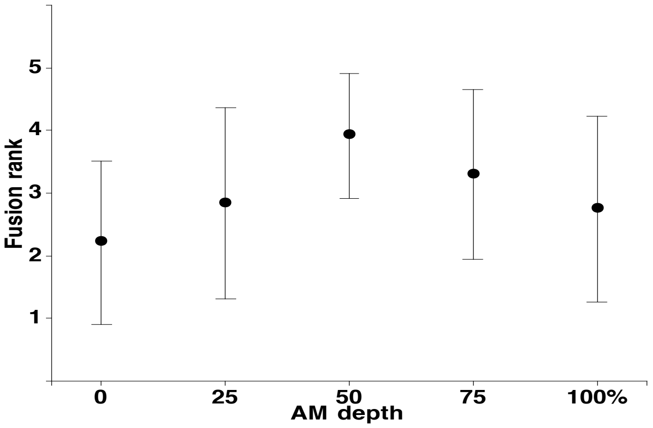


Figure 4.3. Average fusion ranks ± 1 SD, *fd* = 150 Hz, *N* = 12.

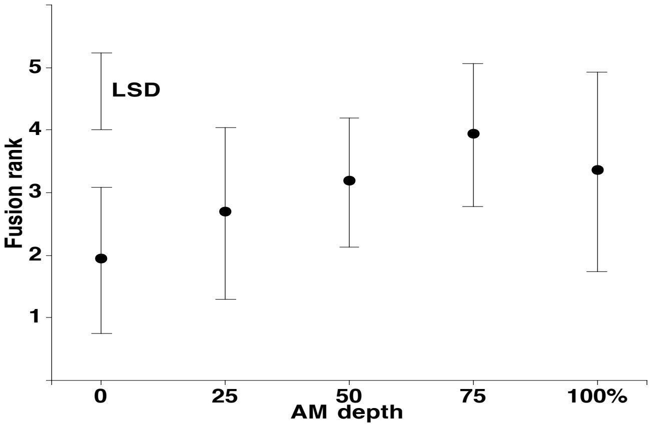


Figure 4.4. Average fusion ranks ± 1 SD, *fd* = 300 Hz, *N* = 12.

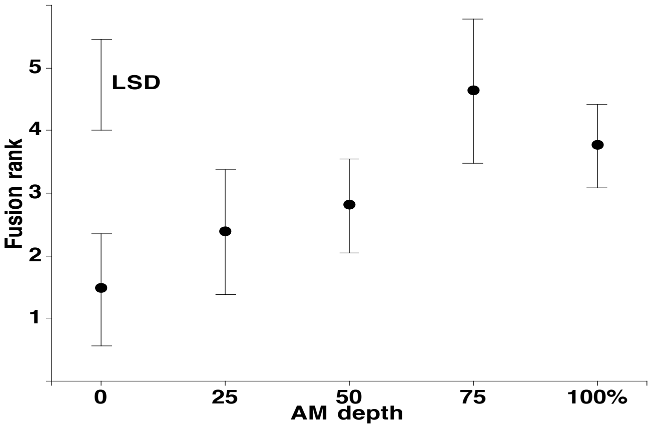


Figure 4.5. Average fusion ranks ± 1 SD, *fd* = 600 Hz, *N* = 12.

A *Friedman test* was used to analyse the subjects' rankings. The Friedman test is a non-parametric test corresponding to a two-way analysis of variance with block factors (Friedman, 1937). A common form of the Friedman statistic is

$$T = b(t - 1)(S_t - C)/(S_r - C),$$

where b is the number of treatments (in this case, the five AM depths), t the number of blocks (the subjects' rankings), S_t the average sum of the squares of the treatments' total rank, S_r the sum of squares of all ranks and $C = {}^1/_4bt(t + 1)^2$ the correction factor for ties. If there are no ties, $S_\mathrm{r} = bt(t + 1)(2t + 1)/6$.

If b and t are not too small, T has approximately a χ^2-distribution with t - 1 degrees of freedom. A better approximation may be

$$T_1 = (b - 1)(S_t - C)/(S_r - S_t)$$

which has approximately an F-distribution with t - 1, (b - 1)(t - 1) degrees of freedom. For details, see Sprent (1993).

The hypotheses to test are H_0: *there is no difference between the* fd_x against H_1*: at least one* fd_x *has a different rank from the others*; fd_x denotes the average rank of fusion at the x% AM depth (x = 0, 25, 50, 75 or 100). The results of the Friedman test applied to the data in Appendix 3 are shown in table 4.1.

Frequency (Hz)	T	T_1
150	7.92	2.17
300	11.15	3.33
600	29.48	17.51

Table 4.1. The results of the Friedman test.

The critical values for significance at the 0.05, 0.01 and 0.001 levels are, respectively, 9.49, 13.28 and 18.47 for the χ^2-distribution with 4 degrees of freedom and 2.58, 3.78 and 5.59 for the F-distribution with 4, 44 degrees of freedom.

Hence, at 150 Hz, there is no significant difference between the fd_x at the 0.05 level. (There is, however, a significant difference at the 0.1 level.)

At 300 Hz, there is a significant difference between the fd_x at the 0.05 level.

At 600 Hz, there is a significant difference between the fd_x at the 0.001 level.

To determine the fd_x that differ from each other, a *least significant difference* (LSD) test was used (Sprent, 1993). The requirement for least significant difference based on the Friedman test is

$$| s_i - s_j | > t_{(b - 1)(t - 1), \alpha}\sqrt{}[2b(S_r - S_t)/(b - 1)(t - 1)],$$

where s_i and s_j are the rank totals for treatments i, j (e. g. fd_0 and fd_{100}) and $t_{(b - 1)(t - 1), \alpha}$ is the t value required for significance for (b - 1)(t - 1) degrees of freedom at the α level of significance. This level must at least be equal to the level indicated by the Friedman test.

LSD tests are to be used only if the treatments to be compared are selected prior to the inspection of the data. In this case, the assumption that fusion should increase with increasing AM depth allows for the comparison of the individual fd_x.

At 150 Hz, the Friedman test did not find a significant difference and, therefore, no LSD were calculated.

At 300 Hz and $\alpha = 0.05$, fd_{50}, fd_{75} and fd_{100} were all significantly higher than fd_0. In addition, fd_{75} was significantly higher than fd_{25}.

At 600 Hz and $\alpha = 0.001$, fd_{75} and fd_{100} were both significantly higher than fd_0. fd_{75} was significantly higher than fd_{50} and fd_{25}.

Hence, there is strong evidence that subjects judge fd_{75} and fd_{100} the highest, but there is not sufficient evidence that fusion increases monotonically with AM depth. While inspection of the data in figs. 4.3-5 suggests that fusion increases from fd_0 to fd_{75} (fd_{50} for the 150 Hz test) and then decreases, no significant differences between fd_{75} and fd_{100} were found.

4.5. Discussion

Clearly, the level of significance increases with the frequency. There are several possible explanations for this. Firstly, since all subjects started with the 150 Hz test and finished with the 600 Hz test, some learning may have occurred. Although at least three of the subjects (nos. I, II and III) were remarkably consistent in their judgements, the average number of times each stimuli was played decreased from approximately ten on average for 150 Hz to approximately six on average for both 300 and 600 Hz. This may be interpreted either as an indication of learning or as easier detection at higher frequencies. The former seems more likely. In an experiment better designed to account for learning, the stimuli should have been presented in different orders to different groups of subjects, e. g. (150, 300, 600 Hz), (300, 600, 150 Hz) and (600, 150, 300 Hz) for three groups of four subjects.[12]

Secondly, fusion may be constant over a fairly large frequency range or the components may fuse more easily at higher frequencies. If this is the case, fusion can not be a result solely of phase synchrony, but AM detection may still play a role even at 600 Hz. Indeed, some of the subjects remarked that the task was easier at 300 Hz, which coincides with the reported AM detectability maximum.

Thirdly, the subjects may have been ranking not the noise-tone fusion, but a different phenomenon, e. g. masking, noise level or some other type of timbre change. It is very difficult to determine whether this is the case. This is a critique that could also be raised against the research of Hermes (1991) or Fales and McAdams (1994). For further discussions, see these papers. In the present experiment, the subjective adjustment of the noise level may not have been sufficient to remove all differences between the stimuli.

Since the stimulus frequencies varied, it is possible to check if the subjects ranked the stimuli according to pitch. The minimum interval was 80:79 or 22 cents and clearly audible, the maximum 8:7 or 231 cents, somewhat larger than a whole tone. The data, shown in Appendix 4, appear almost random, but show a very weak trend towards a higher ranking of higher frequencies. Since the frequencies were randomly selected, there are unfortunately not equal number of data points at the different frequencies. To provide a sounder statistic, the stimulus frequencies should not have been randomised, but presented in different orders to different groups of subjects.

A reasonable conclusion is that the differences in fusion between different AM depths are not harder to detect at 150 Hz than at 600 Hz. Hence, noise-tone fusion is probably constant over a larger frequency range than previously believed. In spite of the fact that few significant differences were found between successive fd_x, I believe it is safe to conclude that increasing AM depth yields higher fusion. Although somewhat speculative, there may also be an explanation for the low rank of fd_{100}: since the stimuli sound more like a flute or a flue pipe than other wind instruments (including voice), subjects may have preferred the more continuous noise of the 75% AM depth.

[12] This was pointed out to me by Gunnar Englund (2001), but unfortunately after the completion of the experiment.

5. Conclusions and future research

If we accept the conclusion presented above, we have to question the hypothesis that detection of phase synchronicity is the main cause of noise-tone fusion, and since the pitch sensation of SAM disappear with modulation frequencies above 500 Hz, the detection of SAM pitch cannot be a cause of fusion at frequencies above 1000 Hz, but it may still work at 600 Hz. To determine the upper frequency limit of fusion, more research is needed.

However, as the research on pitch in IRN shows, there may be other temporal cues, not based on envelope detection or phase synchrony but perhaps on autocorrelation. Some of the mechanisms responsible for fusion are probably to be sought at the low-level features of the auditory system, but this has to be examined by researchers more competent in the field of neurosciences. In the following, I shall discuss some other questions raised by the present research.

Can the auditory system distinguish between simultaneous noise sources amplitude modulated with different frequencies? This seems to be a considerably harder task than separating two harmonic tones, especially since we know that noise is sometimes used to make instruments blend better (see section 2.3), but we know that detection of random noise in IRN is easier than detection of random noise in random noise (Yost, Patterson and Sheft, 1998).

An implicit assumption of this thesis and most other research is that Gaussian white noise sounds better—or more ‘natural’—than e. g. rectangularly distributed noise. Dubnov, Tishby and Cohen (1997) used higher-order statistics (HOS) (or polyspectra) in the analysis of jitter in musical instruments, and showed that there is a correspondence between HOS, e. g. skewness and kurtosis, and the classification of acoustic instruments into families like strings and woodwinds. Are listeners also able to distinguish between different probability distributions and different HOS of noise?

Finally, an interesting research topic would be to compare the analysis methods discussed in section 2.6 with respect to their handling of noise and transients.

Acknowledgements

I would like to thank to my supervisor Sten Ternström and professor Johan Sundberg for their patience and kind support. Thanks also to all the subjects who willingly participated in the listening test, to Gunnar Englund for scrutinising my data analysis and to EMS (*Electroacoustic Music in Sweden*) for providing me with workspace. And many thanks to Love Mangs for his loyal friendship.

References

Bregman, A. S. **1990**. *Auditory scene analysis*. MIT Press, Cambridge (USA).

Burns, E. M. and Viemeister, N. F. **1981**. "Played-again SAM: Further observations on the pitch of amplitude-modulated noise." *Journal of the Acoustical Society of America*, Vol. 70, No. 6, pp. 1655-1660.

Carlson, R., Granström, B. and Karlsson, I. **1991**. "Experiments with voice modelling in speech synthesis." *Speech Communication*, Vol. 10, No. 5-6, Dec. 1991, pp. 481-489.

Chafe, C. **1990**. "Pulsed noise in self-sustained oscillations of musical instruments." *ICASSP 90. Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 1157-1160. IEEE, New York.

Chafe, C. **1992**. "Musical synthesizer system and method using pulsed noise for simulating the noise component of musical tones." United States Patent No. 5,157,216.

Chafe, C. **1993**. "Adding pulsed noise to a flute physical model." ASA 126th Meeting, Denver. Abstract at `http://www.auditory.org/asamtgs/asa93dnv/4aMU/4aMU3.html`.

Chafe, C. **1996**. "Music synthesizer and method for simulating period synchronous noise associated with air flows in wind instruments." United States Patent No. 5,508,473.

Dolson, M. **1986**. "The Phase Vocoder: A Tutorial." *Computer Music Journal*, Vol. 10, No. 4 (Winter 1986), pp. 14-27.

Dubnov, S., Tishby, N. and Cohen, D. **1997**. "Polyspectra as Measures of Sound Texture and Timbre." *Journal of New Music Research*, Vol. 26, No. 4 (December 1997), pp. 277-314.

Englund, G. **2001**. Private communication.

Fales, C. and McAdams, S. **1994**. "The Fusion and Layering of Noise and Tone: Implications for Timbre in African Instruments." *Leonardo Music Journal*, Vol. 4, pp. 69-77.

Friedman, M. **1937**. "The use of ranks to avoid the assumptions of normality implicit in the analysis of variance." *Journal of the American Statistical Association*, Vol. 32, pp. 675-701.

Goodwin, M. **1996**. "Residual Modeling in Music Analysis-Synthesis." *ICASSP 96. Proceedings of the 1996 International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 1005-1008. IEEE, New York.

Hermes, D. J. **1991**. "Synthesis of breathy vowels: some research methods." *Speech Communication*, Vol. 10, No. 5-6 (December 1991), pp. 497-502.

Hirschberg, A. and Verge, M. P. **1995**. "Turbulence Noise in Flue Instruments." *ISMA 95*. IRCAM, Paris.

Laroche, J. and Dolson M. **1999**. "Improved Phase Vocoder Time-Scale Modification of Audio." *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 3, pp. 323-332.

Masri, P. **1996**. *Computer Modelling of Sound for Transformation and Synthesis of Musical Signals*. Ph. D. dissertation, University of Bristol.

Masri, P. and Bateman, A. **1996**. "Improved Modelling of Attack Transients in Music Analysis-Resynthesis." In *Proceedings of the International Computer Music Conference (ICMC96)*, pp. 100-103. The International Computer Music Association, Hong Kong.

Masri, P., Bateman, A. and Canagarajah, N. **1997 a**. "A Review of Time-Frequency Representations, with Application to Sound/Music Analysis-Resynthesis." *Organised Sound*, Vol. 2, No. 3, pp. 193-205.

Masri, P., Bateman, A. and Canagarajah, N. **1997 b**. "The Importance of the Time-Frequency Representation for Sound/Music Analysis-Resynthesis." *Organised Sound*, Vol. 2, No. 3, pp. 207-214.

McAdams, S. **1984**. *Spectral fusion, spectral parsing and the formation of auditory images*. Ph. D. dissertation, Stanford University.

McAngus Todd, N. P. and Lee, C. S. **1998**. “A Sensory-Motor Theory of Rhythm, Time Perception and Beat Induction.” (Draft. Published in a slightly different version in *Journal of New Music Research*, Vol. 28, No. 1.)

McAulay, R. J. and Quatieri, T. F. **1986**. “Speech Analysis/Synthesis based on a Sinusoidal Representation.” *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 34, No. 4, pp. 744-754.

McIntyre, M.E., Schumacher, R.T. and Woodhouse, J. **1981**. “Aperiodicity in Bowed-String Motion.” *Acustica*, Vol. 49, pp. 13-32.

Menell, P., McAnally, K. I. and Stein, J. F. **1999**. “Psychophysical Sensitivity and Physiological Response to Amplitude Modulation in Adult Dyslexic Listeners.” *Journal of Speech, Language and Hearing Research*, Vol. 42, pp. 797-803.

Moore, B. C. J. **1997**. *An Introduction to the Psychology of Hearing*. Academic Press, San Diego.

Nketia, J. H. K. **1974**. *The Music of Africa*. Norton, New York.

Richard, G., d’Allesandro, C. and Grau, S. **1993**. “Musical noises synthesis using random Formant Waveforms.” *SMAC 93. Proceedings of the 1993 Stockholm Music Acoustics Conference*, pp. 580-583. Royal Institute of Technology, Stockholm.

Rocchesso, D. and Turra, F. **1993**. “A generalized excitation for real-time sound synthesis by physical models.” *SMAC 93. Proceedings of the 1993 Stockholm Music Acoustics Conference*, pp. 584-588. Royal Institute of Technology, Stockholm.

Rodet, X. **1980**. “Time-Domain Formant-Wave-Function Synthesis.” In Simon, J. C. (ed.). *Spoken Language Generation and Understanding*. D. Reidel, Dordrecht. Reprinted in *Computer Music Journal*, Vol. 8, No. 3 (Fall 1984), pp. 9-14.

Sandell, G. **2000**. `http://sparky.parmly.luc.edu/sandell/homepage/publications.html`.

Schaeffer, P. **1966**. *Traité des Objets Muscaux*. Seuil, Paris.

Schumacher, R. T. **1993**. “Aperiodicity, subharmonics and noise in bowed string instruments.” *SMAC 93. Proceedings of the 1993 Stockholm Music Acoustics Conference*, pp. 428-431. Royal Institute of Technology, Stockholm.

Serra, X. and Smith, J. O. **1990**. “Spectral Modeling Synthesis: A Sound Analysis/Synthesis System based on Deterministic plus Stochastic Decomposition.” *Computer Music Journal*, Vol. 14, No. 4 (Winter 1990), pp. 12-24.

Serra, X. **1997**. “Musical Sound Modeling with Sinusoids plus Noise”. In Roads, C. et al. *Musical Signal Processing*. Swets & Zeitlinger, Lisse.

Skoglund, J. and Kleijn, W. B. **1998**. “On Time-Frequency Masking in Voiced Speech.” In Skoglund, J. *From Modeling to Perception – Topics in Speech Coding*. Ph. D. dissertation, Chalmers University of Technology, Göteborg.

Smalley, D. **1986**. “Spectro-morphology and Structuring Processes.” In Emmerson, S. (ed.). *The Language of Electroacoustic Music*. London, Macmillan.

Sprent, P. **1993**. *Applied Nonparametric Statistical Methods. Second edition*. Chapman & Hall, London.

Sundberg, J. **1966**. *Mensurens betydelse i öppna labialpipor.* (*The Significance of the Scaling in Open Flue Organ Pipes.*) Ph. D. dissertation, Uppsala University.

Verma, T. S. and Meng, T. H. Y. **2000**. “Extending Spectral Modeling Synthesis with Transient Modeling Synthesis.” *Computer Music Journal*, Vol. 24, No. 2 (Summer 2000), pp. 47-59.

Yost, W. A., Patterson, R. and Sheft, S. **1998**. “The role of the envelope in processing iterated rippled noise.” *Journal of the Acoustical Society of America*, Vol. 104, No. 4, pp. 2349-2361.

Appendix 1. Aladdin source file listing

```
program : Interactive DSP Workbench , version=1.30.03 ;
window : pos=6:1:568:332:1, class=sheet;
model : size=10:18, file=am-final.ald, date=, freq=16000, numfreq=16000, options=2049,
          errors=65535 ;
object binop18 : pos=0:1, class=binop:tap, input=unop3 ;
object amp10 : pos=0:2, class=amp, update=event1, gain=amp10gain ;
object unop3 : pos=0:4, class=unop:load, data=1 ;
object unop5 : pos=0:5, class=unop:load, data=0.001 ;
object amp6 : pos=0:6, class=amp, update=event1, gain=amp6gain ;
object binop7 : pos=0:7, class=binop:sum, input=unop3 ;
object binop13 : pos=0:9, class=binop:tap, input=unop10 ;
object clipper2 : pos=0:10, class=clipper:floor, data=0 ;
object unop13 : pos=0:11, class=unop:neg, data=0 ;
object amp5 : pos=0:12, class=amp, update=reset, gain=amp5gain ;
object binop19 : pos=0:13, class=binop:sum, input=amp10 ;
object sine1 : pos=1:3, class=sine, update=event1, freq=amp2gain, phase=sine1phase ;
object binop9 : pos=1:4, class=binop:sum, input=binop7 ;
object unop7 : pos=1:5, class=unop:log, data=1 ;
object fpfilter4 : pos=1:6, class=fpfilter:hp, update=event1, freq=fpfilter4freq ;
object clipper3 : pos=1:7, class=clipper:floor, data=-4 ;
object fpfilter3 : pos=1:8, class=fpfilter:lp, update=event1, freq=fpfilter3freq ;
object unop10 : pos=1:9, class=unop:mul, data=0.2 ;
object binop3 : pos=1:11, class=binop:sum, input=delaytap3 ;
object binop21 : pos=1:12, class=binop:prod, input=fpfilter1 ;
object sampler1 : pos=1:13, class=sampler:4609, handler=dataduct5, ratio=1 ;
object sampler5 : pos=1:14, class=sampler:515, handler=fft1, ratio=1 ;
object da0 : pos=1:16, class=da:0, update=event1, gain=da0gain ;
object da1 : pos=1:17, class=da:1, update=event1, gain=da0gain ;
object fft1 : pos=2:10, class=fft:26113, size=256, update=sampler5, window=blackman,
          log=true, ratio=10, queue=2 ;
object dataduct5 : pos=2:12, class=dataduct:30208, update=sampler1, size=256, queue=9:2 ;
object noise1 : pos=3:1, class=noise, update=reset ;
object binop5 : pos=3:2, class=binop:diff, input=binop16 ;
object binop2 : pos=3:3, class=binop:prod, input=binop19 ;
object formant2 : pos=3:5, class=formant:pole, update=reset,
          freq=formant2freq, bandw=formant2bandw ;
object fpfilter7 : pos=3:6, class=fpfilter:hp, update=event1, freq=formant2freq ;
object amp3 : pos=3:7, class=amp, update=reset, gain=amp3gain ;
object delayfix2 : pos=3:10, class=delayfix, delay=1000 ;
object delaytap3 : pos=3:11, class=delaytap:delay, update=reset, source=delayfix2,
          delay=delaytap3delay ;
object sampler2 : pos=3:13, class=sampler:37378, handler=dataduct9, ratio=1,
          input=unop10 ;
object noise2 : pos=4:1, class=noise, update=reset ;
object binop16 : pos=4:2, class=binop:sum, input=noise5 ;
object dataduct9 : pos=4:9, class=dataduct:30208, update=sampler2, size=256, queue=11:2 ;
object noise5 : pos=5:1, class=noise, update=reset, data=987610 ;
object counter1 : pos=5:2, class=counter, update=reset, freq=counter1freq ;
object event1 : pos=5:3, class=event, variant=zero, id=5, handler=main ;
object mux1 : pos=7:13, class=mux:16384, hold=250, update=event1,
          list=1.0: 0.0: 0.0: 0.0 ;
object fpfilter1 : pos=7:14, class=fpfilter:lp, update=reset, freq=fpfilter1freq ;
object sine1phase : class=inparam, source=constant, default=0, scale=1, offset=0,
          log=false, zero=false, fix=false, min=-360, max=360, midictl=0 ;
object amp1gain : class=inparam, source=constant, default=50, scale=1, offset=-60,
          log=true, zero=true, fix=false, min=0, max=60, midictl=0 ;
object da0gain : class=inparam, source=constant, default=53, scale=1, offset=-60,
          log=true, zero=true, fix=false, min=0, max=60, midictl=0 ;
object formant2freq : class=inparam, source=constant, default=4000, scale=1, offset=0,
          log=false, zero=false, fix=false, min=0, max=8000, midictl=0 ;
object formant2bandw : class=inparam, source=constant, default=4000, scale=1, offset=0,
          log=false, zero=false, fix=false, min=50, max=6000, midictl=0 ;
object amp3gain : class=inparam, source=constant, default=20, scale=1, offset=-60,
          log=true, zero=true, fix=false, min=0, max=60, midictl=0, input=amp5gain ;
```

```
object amp2gain : class=inparam, source=constant, default=150, scale=1, offset=0,
           log=false, zero=false, fix=false, min=0, max=700, midictl=0 ;
object counter1freq : class=inparam, source=constant, default=100, scale=1, offset=0,
           log=false, zero=false, fix=false, min=0, max=2000, midictl=0 ;
object fpfilter4freq : class=inparam, source=constant, default=100, scale=1, offset=0,
           log=false, zero=false, fix=false, min=0, max=8000, midictl=0,
           input=formant2freq ;
object amp6gain : class=inparam, source=constant, default=50, scale=1, offset=0,
           log=false, zero=false, fix=false, min=0, max=500, midictl=0 ;
object formant3freq : class=inparam, source=constant, default=6, scale=1, offset=0,
           log=false, zero=false, fix=false, min=0, max=20, midictl=0 ;
object formant3bandw : class=inparam, source=constant, default=14, scale=1, offset=0,
           log=false, zero=false, fix=false, min=1, max=15, midictl=0 ;
object amp5gain : class=inparam, source=constant, default=100, scale=0.01, offset=0,
           log=false, zero=false, fix=false, min=0, max=100, midictl=0 ;
object delaytap3delay : class=inparam, source=constant, default=0, scale=0.1, offset=0,
           log=false, zero=false, fix=false, min=0, max=100, midictl=0 ;
object amp10gain : class=inparam, source=signal, default=50, scale=-0.2, offset=-1,
           log=false, zero=false, fix=false, min=0, max=60, midictl=0, input=amp5gain ;
object amp8gain : class=inparam, source=constant, default=1, scale=1, offset=0,
           log=false, zero=false, fix=false, min=0, max=1, midictl=0 ;
object fpfilter1freq : class=inparam, source=constant, default=5, scale=1, offset=0,
           log=false, zero=false, fix=false, min=0, max=2000, midictl=0 ;
object fpfilter3freq : class=inparam, source=constant, default=6000, scale=1, offset=0,
           log=false, zero=false, fix=false, min=0, max=8000, midictl=0,
           input=formant2freq ;
window : pos=570:3:826:164:1, class=spectrum, data=FFT1, size=256, channels=1;
window : pos=73:406:328:469:1, class=slider, inparam=amp3gain;
window : pos=79:472:323:677:1, class=xymap, x=formant2freq, y=formant2bandw;
window : pos=3:333:1023:404:1, class=slider, inparam=amp2gain;
window : pos=829:80:941:261:1, class=slider, inparam=da0gain;
window : pos=667:405:923:467:1, class=slider, inparam=amp6gain;
window : pos=342:470:675:679:1, class=waveform, data=dataduct9, size=256, channels=2;
window : pos=696:477:808:675:1, class=slider, inparam=amp5gain;
window : pos=368:405:624:468:1, class=slider, inparam=delaytap3delay;
window : pos=575:170:822:288:1, class=waveform, data=dataduct5, size=512, channels=1;
```

Appendix 2. Test instructions

Test instructions in Swedish

Ljuden nedan består av en ton samt brus.

Rangordna ljuden efter hur bra du tycker att tonen och bruset hör ihop (”fusion”). Ju bättre de hör ihop, desto högre upp skall de placeras på skalan till vänster.

Det viktiga är ordningen mellan ljuden, inte avståndet mellan dem eller exakt var på skärmen de placeras. Det spelar ingen roll hur långt till vänster eller höger de placeras.

Om du inte hör någon skillnad, går det bra att placera ljuden bredvid varandra.

Dubbelklicka på ljuden för att spela dem. Du kan lyssna på ljuden flera gånger. Tryck Esc-tangenten om du vill avbryta uppspelningen.

När du är nöjd med din rangordning, tryck ”Finished!”.

English translation

Each sound below consists of a tone and noise.

Please rank the sounds according to how well the tone and noise “fuse” together. Better fusion yields a higher placement on the scale to the left.

The only thing that is important is the order of the sounds, not the distances between them or their exact position on the screen. It does not matter how far to the left or right they are placed.

If you do not perceive any difference, you should place the sounds at the same level.

Double-click on the sounds to play them. You may listen to the sounds several times. Press the Escape key if you want to stop the playing.

When you are satisfied with your ranking, press “Finished!”.

Appendix 3. Fusion ranks

AM depth (%)	*Subject* I	*Subject* II	*Subject* III	*Subject* IV	*Subject* V	*Subject* VI	*Subject* VII	*Subject* VIII	*Subject* IX	*Subject* X	*Subject* XI	*Subject* XII
0	1	1	2	2	1	4	1,5	2	3	1	3	5
25	2	2	1	4	5	5	3	1	4	2	1	4
50	3	3	3	3	4	3	5	5	5	5	5	3
75	5	5	4	5	3	2	1,5	4	2	4	2	2
100	4	4	5	1	2	1	4	3	1	3	4	1

Table 1. Fusion ranks, *fd* = 150 Hz.

AM depth (%)	*Subject* I	*Subject* II	*Subject* III	*Subject* IV	*Subject* V	*Subject* VI	*Subject* VII	*Subject* VIII	*Subject* IX	*Subject* X	*Subject* XI	*Subject* XII
0	1	1	1	1	2	4	2	2	3	1	1	4
25	2	2	2	3	1	5	5	1	4	2	2	3
50	3	3	3	4	4	3	1	3	5	3	4	2
75	4,5	5	5	5	3	2	3	4	2	4,5	4	5
100	4,5	4	4	2	5	1	4	5	1	4,5	4	1

Table 2. Fusion ranks, *fd* = 300 Hz.

AM depth (%)	*Subject* I	*Subject* II	*Subject* III	*Subject* IV	*Subject* V	*Subject* VI	*Subject* VII	*Subject* VIII	*Subject* IX	*Subject* X	*Subject* XI	*Subject* XII
0	2	1	1	1	1	1	2	1	1	1,5	1	4
25	1	2	2	2,5	2	3	5	2	2	1,5	2,5	3
50	3	3	3	2,5	3,5	2	4	3	3	3	2,5	1
75	5	5	5	5	5	5	1	5	5	4,5	5	5
100	4	4	4	4	3,5	4	3	4	4	4,5	4	2

Table 3. Fusion ranks, *fd* = 600 Hz.

Appendix 4. Fusion ranks and stimulus frequencies

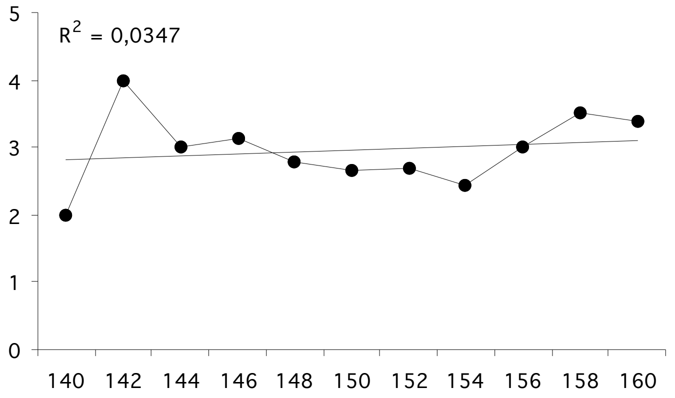


Figure 1. Relation between fusion rank and stimulus frequency, 150 Hz.

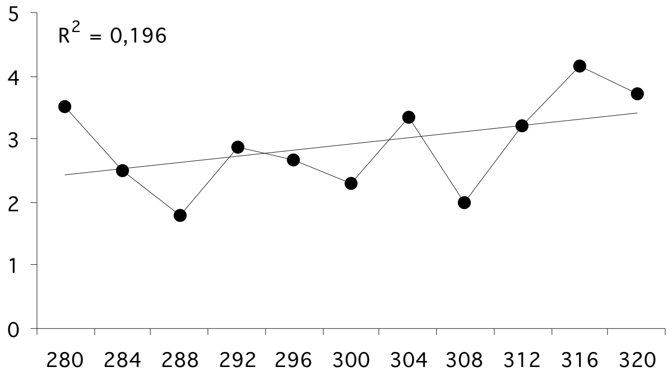


Figure 2. Relation between fusion rank and stimulus frequency, 300 Hz.

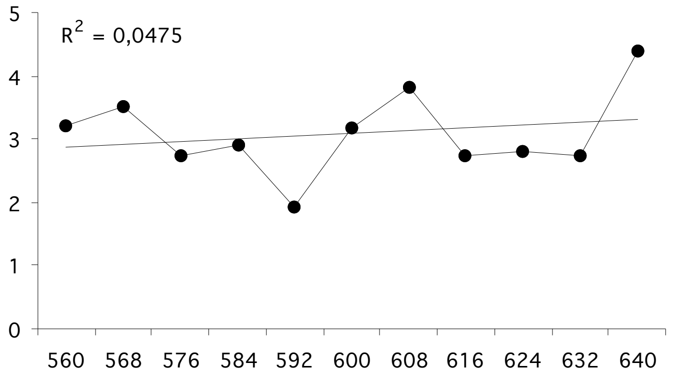


Figure 3. Relation between fusion rank and stimulus frequency, 600 Hz.