PERCEPTUAL FEATURES OF SOME CONSONANTS STUDIED IN NOISE

BY

MARION RÉGNIER

B.S., Ecole Superieure d’Electricite, France, 2005

THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Electrical and Computer Engineering in the Graduate College of the University of Illinois at Urbana-Champaign, 2007

Urbana, Illinois
To France Lopez.
I would like to thank my adviser, Dr. Jont Allen, for opening my eyes to the very important topic of human speech recognition and its field of applications. I am very proud to have had the opportunity to work with him and study this topic for my thesis. Thank you to the other members of my research group: Feipeng Li, Bryce Lobdell, Andy Lovitt, Sandeep Phatak, and Nadeeka Yapa for their very useful input and team spirit.

I would like to thank my parents, Jérôme and Michèle Régnier, for their continuous support; my sister, Margaux, and my two brothers, Romain and Raphaël. Also, thanks to my close friends, Jérôme Autric, and Solene Roure. In addition, I would like to thank all my friends from UIUC, both students and professors.

Lastly, I thank the Ecole Supérieure d’Electricité - Supélec, for providing me with the opportunity to study abroad at the University of Illinois.
# TABLE OF CONTENTS

## CHAPTER 1  INTRODUCTION  
1.1 Overview  
1.2 Motivation  
1.3 Literature Review  
1.3.1 Irrelevance of language and context effects  
1.3.2 Confusion matrices  
1.3.3 Articulation index  
1.4 Coarticulation  
1.4.1 Coarticulation: An ill-defined concept  
1.4.2 A metric for coarticulation  
1.4.3 Nasal murmur and vocalic transition  
1.5 UIUC-S04 Experiment  
1.6 Perceptual Models  
1.6.1 The AI-gram  
1.6.2 The short-time AI  
1.6.3 The event-gram  
1.7 The Short-Time Fourier Transform  

## CHAPTER 2  PILOT STUDIES  
2.1 Four-Step Analysis for /tɛ/ Utterances  
2.2 Truncation Experiment  
2.3 /ma/-/na/ Discrimination  

## CHAPTER 3  METHODS  
3.1 Overview  
3.2 My Language Background  
3.3 Subject Selection  
3.4 Hardware and Software  
3.5 Talkers and Utterances  
3.5.1 /m/ and /n/  
3.5.2 /p/, /t/, and /k/  
3.5.3 List of utterances  
3.6 Noise and SNR  
3.7 Modification of the Speech Files  
3.8 Experiment Structure  

v
CHAPTER 1

INTRODUCTION

1.1 Overview

The purpose of this thesis is to find the perceptual features used by normal hearing listeners to identify consonants in noise. Distinctive features, or production features have traditionally been used to analyze how humans decode speech. However, those features have little to do with the robustness of consonants in noise. An experiment where specific parts of the speech are removed, given that these parts have been identified as carriers of the robust-to-noise features, called events, would reveal part of the acoustic to perceptual map and demonstrate that consonants share common cues. This research takes us a step forward toward the design of new processing algorithms to improve hearing aids, cochlear implants, and automatic speech recognition machines.

This document is organized as follows. This chapter will take the reader through the motivation behind this research, a literature review, an analysis of previous experiments results that motivated this research, and an introduction to the various tools and methods used. Chapter 2 consists of the preliminary analysis carried out to identify certain events, later on tested in the Verification experiment described in this thesis. Chapter 3 summarizes the experimental protocol for the Verification experiment conducted. Results are reported in Chapter 4. Chapter 5 discusses the results and analysis in greater depth. Future extensions of this work and its conclusions are stated in Chapter 6.
1.2 Motivation

After 50 years of work, a basic understanding of the speech code remains a mystery. Such a theory of speech perception is critical for the development of new hearing aids and cochlear implants, and for speech recognition. To research this long-standing problem, it is necessary to collect listeners' responses to syllables in noise and correlate their confusions with the utterances acoustic cues. Such confusions made are far from random. If we could manipulate the spectrotemporal features, we could prove the existence of perceptual cues, called *events*, that are used by listeners to discriminate consonants the intelligibility of the modified consonants. Such processing could well lead to a new family of hearing aids as well as robust automatic speech recognition. Having an automatic speech recognition (ASR) device based on human speech recognition would be a tremendous breakthrough to make speech recognizers robust to noise, an area where significant improvement is needed.

The approach I am following in this research aims at correlating the acoustic information present in the noisy speech to human listeners responses to the sounds, where human communication can be interpreted as an “Information channel” in the Shannon sense [1], as shown in Fig. 1.1. We are working on the receiver (RX) side, and trying to identify the variables that the ear is most sensitive to when listening in noisy environments. First I carry out an analysis of previous experiments in order to identify the robust-to-noise cues, then the experimental portion of this thesis tested that the listeners were either sensitive to the removal of those cues or insensitive to certain parts of the speech. In the latter case, their recognition would not be impaired. The implications of such characterization are multiple. For example, instead of amplifying both speech and noise, tuning a hearing aid to extract these cues and amplify them on a listener basis could result in a great improvement of speech identification in noisy environments.
1.3 Literature Review

1.3.1 Irrelevance of language and context effects

One might wonder why we study phonology (consonant-vowel sounds, noted CVs) rather than language (context). It is frequently stated, and seems to be widely believed, that in order to model natural speech understanding, it is necessary to include the context effects of phonology, phonotactics, morphology, and lexical constraints, and that a complete understanding of speech perception will also require syntactic and semantic context constraints. However, humans are capable of accurate phonetic transcription under high noise conditions, with only phonological constraints (they can accurately transcribe meaningless CV sounds in large amounts of noise [2]). Machine speech recognition of simple sounds is comparatively much worse [3].

It follows that a gap exists in our basic understanding of simple CV speech perception. As first shown in detail by Fletcher, and in more detail by Boothroyd et al., the average phone score does an excellent job of characterizing the sentence intelligibility score, having significant context, with a variance due to experimental observation [4,5]. Thus the impact of context may be modeled with simple single-parameter models, as a function of the
average phone score \( s \), with small variance \([6]\).

The development of information theory by Shannon in 1948 raised the bar \([1]\). Using information theory, Miller \([7]\) and Miller and Isard \([8]\) also showed such relationships, with and without a grammatical context. Bronkhorst et al. \([9]\) went much further, and accurately simulated human performance using word context, given nonsense phone scores as the input. A detailed analysis of this literature is provided by Allen in \([10]\).

While context effects are critical when decoding natural language, they have little to do with the robust speech code at 0 dB Signal-to-noise ratio (SNR) and below. Such noise robustness has been a major area of misunderstanding and heated debate. This evidence is clear from an analysis of the confusion matrices of CV sounds, measured by Miller and Nicely \([11]\), as emphasized by Allen \([12]\) and Phatak and Allen \([2]\).

### 1.3.2 Confusion matrices

In 1955, G. Miller and P. Nicely (MN55) collected confusion matrix data from from 5 female listeners who also served as talkers, each speaking 50 utterances of 16 CV sounds \( (16 \text{ consonants and vowel } /a/\) under 6 wide-band SNR conditions in white noise. Confusion matrices (CMs) were derived from the results, where a row corresponded to the average spoken CV and columns to the responses. The number of hits define each intersection; therefore, diagonals represent the correct identification and off-diagonal confusions. Such confusion tables can be derived at each SNR yielding a 3D confusion matrix. They averaged the data on a CV basis without analyzing utterance or talker variability. While their analysis represented a major step forward in the understanding of CV confusions, many important questions remained unanswered: What are the acoustic features that explain the confusion patterns? How do listeners vary in their listening strategy? To what extent are the different consonant confusion patterns affected by utterance
variability? How would their results be modified if they had used more listeners, less trained, who didn’t serve as talkers, and who didn’t know each other?

Since MN55’s results were analyzed in terms of production features such as voicing, place, or manner, little is known about the spectrotemporal information present in each waveform that could have explained the different confusions. To gain access to the missing utterance waveforms for subsequent analysis and further explore the unknown effects of the noise spectrum, the UIUC Human Speech Recognition (HSR) Group at the Beckman Institute conducted two listening experiments, denoted MN05 [13, 14] and UIUC-S04 [2]. The latter will be briefly described in the next section.

The classic study of Wang and Bilger in 1973 [15] pointed out the ineffectiveness of a then popular analysis method (MDS), multidimensional scaling, in confusion matrices analysis, [10,12,16,17]. Unfortunately, Wang and Bilger averaged their CMs across SNR, greatly reducing the utility of their data. They also studied confusions in hearing impaired ears but with no masking noise and without further analysis on the hearing impaired’s high sensitivity to noise.

Another classic study, Dubno and Levitt [18], analyzed confusions of a few sounds in normal hearing listeners, but the data available from their study are limited. Gordon-Salant [19] studied confusions in hearing impaired ears, and found that confusions were listener dependent. The study was unable to identify common speech perception patterns.

Other than the above publications, we are not aware of additional published studies that have looked at the detailed across-consonant confusions for either normal or hearing impaired listeners.

Cooper et al. [20] characterized invariant cues by reducing the amount of information available to the ear: they synthesized simplified CVs based on a short noise burst (10 ms, 400 Hz bandwidth), followed by artificial formant
transitions composed of tones. However, no information is provided about
the robustness of their speech samples to masking noise, nor the importance
of the synthesized features relative to other cues present in natural speech.
A sound can only be perceptually characterized by finding the source of its
robustness and confusions, by varying the SNR [21, 22].

The research reported here focuses on correlating the confusion pat-
terns [2, 12], defined as speech sounds CV confusions versus SNR (CM row,
denoted $P_{\text{heard/spoken}}(\text{SNR})$), with the speech audibility information, in or-
der to explain normal hearing listeners confusions and identify the spec-
trotemporal nature of the perceptual features characterizing those sounds.
Our goal is to identify and label such events by extracting the necessary
information from the listeners’ confusions. We shall show that spectrotemp-
oral cues, composed of across-frequency temporal coincidence define these
events. Our observation supports these coincidences as a basic element of
the auditory object formation.

1.3.3 Articulation index

The articulation index ($AI$) is the foundation stone of speech perception,
laid by Fletcher et al. [23–25]. The AI is detailed in [10], and its basic
concept is to compute maximum entropy average phone scores based on the
SNR in independent frequency bands. The AI is the average critical band
signal to noise ratio, in decibels re sensation level [dB-SL], scaled by the
dynamic range of speech (30 dB) [2, 10, 26].

Allen in 2005 [10] showed that the average phone score $P_c(AI)$, for 11 of
the 16 consonants in MN55 experiment, can be modeled as a function of the
AI, the recognition error $e_{\text{min}}$ at $AI = 1$, and the error $e_{\text{chance}} = 1 - 1/16$
at chance performance ($AI = 0$), so that:

$$P_c(AI) = 1 - P_e = 1 - e_{\text{chance}}e_{\text{min}}^{AI}$$  \hspace{1cm} (1.1)
Allen and Phatak et al. extended the AI formula from [24] to account for the peak-to-RMS ratio for the speech $r_k$ in each band [2, 10], yielding Eq. (1.2). Parameter $K = 20$ bands, called articulation bands, have traditionally been used and determined empirically to have equal contribution to articulation. The AI in each band is noted $AI_k$:

$$AI_k = \min\left(\frac{1}{3}\log_{10}(1 + r_k^2snr_k^2), 1\right), \quad (1.2)$$

where $snr_k$ is the SNR (linear) in the $k^{th}$ articulation band.

The total AI is therefore given by:

$$AI = \frac{1}{K}\sum_{k=1}^{K} AI_k \quad (1.3)$$

The Articulation Index is the basis of many standards, and its long history and utility is discussed in length in several papers and books [10, 12, 26–29].

1.4 Coarticulation

1.4.1 Coarticulation: An ill-defined concept

There does not appear to be a one definitive and precise definition of coarticulation. First, coarticulation is a concept defined on production, not perception, and based on the effect of muscles when speaking. It can be either forward (carry-over) or backward (anticipatory). Carry-over coarticulation is due to the fact that the muscles of the vocal tract have to anticipate the next movement to pronounce the following item.

Coarticulation as we observed in CV perception is characterized by the existence of different spectrotemporal regions for each consonant, depending on the following vowel; therefore, the coarticulation we have seen is anticipatory coarticulation.
1.4.2 A metric for coarticulation

The *locus equation* introduced by [30] has been used extensively by phonologists to develop a discrimination measure for place of articulation in stop consonants. Locus equations are straight lines plotting the onsets of $F2$ transitions as a function of the corresponding midvowel nuclei value, as seen in [31]. The slope of the locus equation is assumed to characterize the degree of coarticulation. The slope and $y$ intercepts are used to segregate place, since the slopes of the locus equations only are insufficient to predict stop place categories, as emphasized in [32].

There also exist other ways to measure coarticulation by varying different features (e.g., at location A) and examining the effect on this variation on location B. Coarticulation is still a debated issue, and to my knowledge the optimal metric has not been found. We do not know of any papers that study coarticulation and speech in noise. Cooper et al. [20] seem to have pioneered the study of perceptual coarticulation, but they used synthetic speech in quiet condition. Our approach is, therefore, closer to what humans hear, thanks to the use of natural speech.

1.4.3 Nasal murmur and vocalic transition

This section briefly discusses some of the literature relevant to the /m/-/n/ discrimination and how it applies to the results from the current experiment. The /m/place of articulation is bilabial whereas /n/is alveolar, and different cues seem to be used for place discrimination between these two nasals. They are produced by an occlusion of the vocal tract by the lips or the tongue and the opening of the velar port.

All nasals in English are voiced. When the vocal tract is closed and the nasal passage open, air radiates from the nose, creating the nasal murmur. Since the shape of the pharyngeal-nasal passage does not change with the sound, the murmur is almost the same for all nasals. This murmur is stronger
in amplitude below 500 Hz. Therefore, even if the length of the vocal tract used as a secondary branch varies and is longer for /m/ than for /n/, there is no salient difference in the murmur spectrum. It does seem to be carrying some discrimination information nonetheless, as described in Kurowski and Blumstein [33]. Repp [34], confirmed these findings but noted that place information seemed “somewhat” more salient in the formant transitions region than in the murmur. However, Repp emphasized the fact that formant movements do not seem to be the invariant cue used for the /m/-/n/ distinction, favoring spectral change from the murmur to the vowel onset as a better cue, at least for back vowel contexts such as /ɑ/. This left the door open to possible invariant discriminations cues across vowel contexts, such as short-term neural adaptation. Therefore, the present study investigated transitions from the consonant to the vowel, leaving the nasal murmur intact.

It is also important to note that from the Shannon channel capacity point of view, as shown in Fig. 1.1, the aforementioned features are located on the transmitter (TX) side, since they are production variables. The research described in this thesis aims at finding the relevant variables on the receiver (RX) side, in our case the human ear.

1.5 UIUC-S04 Experiment

This section describes a Miller and Nicely type of experiment run by Sandeep Phatak and other members of the HSR group in 2004, and was the inspiration for the experimental portion of this thesis. I used the results to select the utterances for the experiment described in this thesis (see Chapter 3). The experiment was implemented with Matlab.

The experiment measured normal hearing listeners responses to 64 CV sounds (16 Cs * 4Vs) extracted from a commercially available database (LDC-2005S22). Unlike MN55, the noise was speech-weighted, which ac-
counts for the differences in confusions between the two experiments. The noise versus speech spectrum is shown in Fig. 1.2 and was also used in the verification experiment described in this thesis.

The goal of the verification experiment was to collect CM data from normal hearing listeners in order to later identify the events (or perceptual features), on which humans rely for robust consonant identification in noise.

Confusion patterns (row of the CM) corresponding to a spoken utterance can be plotted as a function of SNR to have a visual representation. The scores can also be averaged on a CV basis, gathering scores for all utterances of a same CV. Examples of confusion patterns plots from UIUC-S04 are shown in Chapter 2.

1.6 Perceptual Models

1.6.1 The AI-gram

The AI-gram, AI(t, X, SNR), is the AI density as a function of time and frequency (or distance X along the basilar membrane), computed from a cochlear model, having human critical bands, followed by a simple model of
the auditory nerve [35]. Figure 1.3 shows the block diagram of how an AI-gram is computed given a noisy speech signal $s(t)$. This is a simple extension of French and Steinberg’s work on the AI in 1947 [24] but the estimation of the SNR is different. The AI-gram, before the calculation of the AI, adds a conversion of the basilar membrane vibration to a neural firing rate via an envelope detector. Starting from a critical band filter bank, the envelope is determined at the output of each filter, representing the average rate of the neural firing pattern across the cochlear output. The mean ($\mu_i(t)$) and the standard deviation ($\sigma_i(t)$) of the slowly, time varying and frequency dependent noise floor are estimated using the EM algorithm. The signal above $\mu_i + \alpha\sigma_i$, with $\alpha$ a fixed parameter, is considered to be the audible speech that contributes to articulation. These audible speech modulations across frequency are stacked vertically to get a spectrotemporal representation in the form of the AI-gram. For more details, see [35]. Therefore, the AI-Gram is a perceptual model, and its output is assumed to be correlated in this research with our psychophysical experiments through the event-gram (see Sec. 1.6.3). Indeed, when a speech signal is audible, its information is visible in different degrees of black on the AI-gram. If follows that all noise, and speech buried in noise, would appear in white, due to the normalization to the noise floor.

![AI-gram block diagram](image)

Figure 1.3: AI-gram block diagram.
1.6.2 The short-time AI

The integration of the AI-gram over frequency or place $X$ gives the AI as a function of time, called the short-time AI, $ai(t)$. This function is displayed below the AI-gram, with conversion to bits per centisecond [bits/cs]. Rather than expressing the AI using the base-10 log dB measure, a base-2 log is used, to convert the spectrum amplitude into bits per second, as used by Shannon [1] when computing the channel capacity. The integral over frequency (sum over channels) is displayed as a function of time, in bits per centisecond [cs]. In mathematical terms, the short-time AI is $ai(t) = \int_X AI(t, X, SNR) dX$.

1.6.3 The event-gram

The event-gram is a “rotation” of the AI-gram, yielding to a representation of the AI as a function of frequency or place, and SNR, at an instant of interest, $t = t^*$. It is defined as $AI(t^*, X, SNR)$. The noise was fixed when computing the event-gram. The event-gram allows one to visually quantify the correlation between the AI and the confusion patterns, in order to demonstrate the existence of events [21, 22]. An example of how the event-gram is used is described in Chapter 2.

1.7 The Short-Time Fourier Transform

The application of the short-time Fourier transform (STFT) is to break down a signal into different bands by moving the window through the signal of interest. It can be described as the output of a filter bank, as defined by:

$$X_n(e^{j\omega k}) = F\{w(n - m)x(m)\} = \sum_{m=1}^{N} x(m)w(n - m)e^{-j\omega km}, \quad (1.4)$$

where $x(n)$ is the signal analyzed, $w(n)$ the window function, and $X_n(e^{j\omega k})$ is the complex STFT evaluated at each time $n$ and frequency $\omega_k$. The window $w(n)$ used in this thesis is a Kaiser window (the length of the window was
256, and $\beta = 10$ to obtain a $-80$ dB sidebands attenuation). The STFT can be used to represent the signal properties such as speech pitch and formants, but in other applications, an inverse synthesis to reconstruct the signal is required. The application I used in my research consists of modifying a spectrotemporal region in the STFT domain, by scaling the speech signal up or down (e.g. by zero) on this region. The resulting signal is then reconstructed using the overlap add (OLA) method, described in Eq.(1.5). The OLA is based on the definition of $X_m(e^{j\omega k})$ as the Fourier transform of $\hat{x}(m) = x(m)w(m - n)$:

$$y(n) = \sum_{m} \sum_{k} X_m(e^{j\omega k})e^{j\omega k n} \tag{1.5}$$

Therefore, to obtain $y(n)$, we overlap the inverse Fourier transform of $X_m(e^{j\omega k})$ for each $m$, meaning we add several $y_m(n)$, defined as in Eq.(1.6)

$$y_m(n) = F^{-1}(x(n)w(m - n)) \tag{1.6}$$

where $m = \frac{N}{4}k$ samples, with $k$ integer, for a window of duration $N$ [36].

Figure 1.4 is an example of the summation of $y_m(n)$ over the frame number $m$. The analysis and resynthesis is valid only if $\sum_{-\infty}^{\infty} w(n - m) = 1$. Otherwise, the signal is amplitude modulated with period $N/4$. This result can be shown using the Poisson summation formula.

The STFT processing is used in a Matlab program, called Beren, written specifically for this purpose, used to synthesize the speech files for the experiment described in this thesis. This tool allows the user to visually select a region of interest in time and frequency on the AI-gram, and to choose the scaling factor that will be used by the STFT. The scaled region will then appear in different levels of green on the AI-gram depending on the scaling. A display of an AI-gram of the modified file is also possible in recent versions of the program, as presented in Chapter 3.
Figure 1.4: Example of overlapping sections of the OLA method. Each window is of duration $N$. Each window is delayed by $m$ samples as compared to the previous window. All windows are zero padded to the total length of the signal. This figure is from [36].
CHAPTER 2

PILOT STUDIES

2.1 Four-Step Analysis for /tɛ/ Utterances

The four-step analysis method is an analysis that uses the perceptual models described above and correlates them to the confusion patterns. It lead to the development of the event-gram, and shows the relation between the AI-gram, the event-gram and the confusion patterns for a given utterance. I used the four-step method to draw conclusions about the /t/ event, but it could be extended to other consonants as well.

Step 1 of our four-step analysis consists of the classification of the similarities and differences in confusion patterns across a large number of utterances, displayed in the bottom right panels of Figs. 2.1 and 2.2. As the noise is raised from quiet, a confusion threshold is reached, defined as the SNR at which significant confusions occur, making up the confusion group of this utterance. For talker 117 /tɛ/ (Fig. 2.1, bottom right panel), it is located at $\approx -6$ dB SNR, and the group is /p/, /t/, /k/, whereas the confusion threshold is $\approx -20$ dB SNR for talker 112 (Fig. 2.2 bottom right panel), which has a weakly defined group (/k/ and /d/) due to its extreme robustness (very low threshold).

Three major observations may be drawn from step 1 (the confusion patterns analysis of Figs. 2.1 and 2.2): (i) morphing, (ii) robustness and confusion patterns variability, and (iii) priming. First, we observe that as the noise is increased for utterance m117te, /t/ morphs to /p/, meaning that the probability of recognizing /t/ drops, while that of /p/ increases, and even
Figure 2.1: Analysis of sound /tɛ/ spoken by male talker 117.

Figure 2.2: Analysis of sound /tɛ/ spoken by male talker 112.
overcomes the target consonant score. As shown on the confusion patterns plot of Fig. 2.1, the recognition of /p/ is maximum ($P_{/p/} = 60\%$) at an SNR of -16 dB, where the score for /t/ is 6%, after the start of decrease (circled in blue). This effect, while at first surprising, is common in our database, and shows that noise can mask an essential feature for the recognition of a sound, leading to a consistent confusion among our subjects. However, such morphing is not ubiquitous. For example /p/ very rarely morphs to /t/, and utterance m112te does not present such a score inversion. Morphing demonstrates that consonants are not uniquely characterized by independent features, but that they share common cues that could be weighted differently in perceptual space. At -16 dB SNR, the sound /k/ is also heard by our listeners, but only 25% of the time. Therefore, it seems very likely that /t/, /p/ and /k/ share common features, and that the /t/ feature is strong for m112te but weak for m117te. We shall show that this conclusion is further supported by the truncation experiments, inspired by Furui [37], discussed in Section 2.2. We hypothesized that confusions take place when the acoustic features defining the primary /t/ event are masked by the noise, and that the remaining cues are part of what perceptually characterizes competitors /p/ and /k/.

A second observation is that confusion patterns and robustness vary dramatically across utterances of a given CV: unlike for talker m117, /te/ spoken by talker m112 does not morph to /p/ or /k/, and its recognition is higher (Fig. 2.2, bottom right panel). For this utterance, /t/ (solid green line) was accurately identified down to -18 dB SNR (circled in blue), and was still well above chance performance (1/16) at -22 dB. Its main competitors /d/ and /k/ have lower score [11], and only appear at -18 dB SNR.

It is clear that these two /te/ sounds are dramatically different. Such a difference may only be established by the addition of masking noise. For hearing impaired subjects with an increased sensitivity to noise (called an
SNR-loss, when an ear needs a larger SNR for the same speech score), one would predict that their score for utterance m112te would typically be higher than that of utterance m117te, at a given SNR.

Finally, listening experiments show that when the scores for consonants of a confusion group are similar, listeners can prime between these phones. Here, they may switch at will between the two or even three consonants. Listeners may have a bias toward one or the other sound, causing the scores differences. For example, for /te/ (utterance 117, Fig. 2.1), the average listener randomly primes between /t/ and /p/ at around -7 dB SNR, whereas they typically have a bias for /p/ below this value, and for /t/ above. The SNR range for which priming takes place is listener dependent; the confusion patterns presented here are being averaged across listeners and, therefore, are representative of an average priming range. It seems very likely that priming occurs when invariant features, shared by consonants of a confusion group, are at the threshold of being audible, and when one distinguishing feature is masked.

These two /te/ are an example of utterance variability, due to the natural inhomogeneity of the sounds, as shown by the analysis of Step 1: two sounds are heard as the same in quiet, but they are heard differently as the noise intensity is increased.

We conclude that phoneme perception categories are complex, diverse, and categorical. Indeed, the studied utterances have different confusion thresholds, different confusion groups, morphs or not, and may be subject to priming in some SNR range. Thus the hypothesis is that the acoustic representations of the perceptual features are not invariant, but that the perceptual features themselves (events) remain invariant. The unknown mapping from acoustics to event space is what we are exploring in our research. Therefore, we must quantitatively relate the confusions patterns and robustness, analyzed in the above discussion, to the audible cues at a given
SNR, to identify the acoustic features that map to the “perceptual space.” We will demonstrate that these events are common across utterances of a particular consonant, whereas the acoustic correlates of the events highly vary on an utterance basis.

Step 2 of our four-step analysis consists of using the AI-gram [35], described previously in Chapter 1, to analyze the audible speech information at a given SNR. For talker 117, Fig. 2.1 (top left panel) at 0 dB SNR, we observe that the high-frequency burst, having a sharp energy onset, stretches from 2.8 kHz to 7.4 kHz, and runs in time from 16-18 cs (a duration of 20 ms). According to the confusion patterns previously discussed (Fig. 2.1, bottom right panel), at 0 dB SNR consonant /t/ is recognized 88% of the time. The burst for talker 112 has higher intensity and spreads from 3 kHz up, as shown of the AI-gram for this utterance (Fig. 2.2, top left panel), which results in a 100% recognition above -10 dB SNR. This observation leads us to our Step 3, the integration of the AI-gram over frequency, (Bottom right panels of Figs. 2.1 and 2.2). One obtains a representation of the average audible speech information over frequency as a function of time, denoted the short-time AI in bits/cs, ai(t), because the AI is the area under the resulting curve (see methods section). The first maximum, ai(t*) (vertical dashed red line on the top and bottom left panels of Figs. 2.1 and 2.2), is a gross indicator of the audibility of the consonant. The frequency content has been collapsed, and t* indicates the time of the relevant perceptual information for /t/.

The identification of t* allows Step 4 of our correlation analysis. The top right panels of Figs. 2.1 and 2.2 represent the event-grams for the two utterances. As described previously, the event-gram, AI(t*, X, SNR), is defined as a cochlear place (or frequency) versus SNR slice at one instant of time. The event-gram is the link between the confusion patterns and the AI-gram. Unlike the AI-gram, the event-gram displays the speech spectrum
information as a function of SNR, at a given time \( t^* \), previously determined in Step 3. If several AI-grams were stacked on top of each other, at different SNR s, the event-gram can be seen as a vertical slice through this stack. Namely, the event-grams displayed in the top right panels of Figs. 2.1 and 2.1 are plotted at \( t^* \), characteristic of the /t/ burst. Note the horizontal red dashed line line, going from the bottom of the burst on the AI-gram, onto the bottom of the burst on the event-gram at SNR = 0 dB (vertical red dashed line on the event-gram), establishing a visual link between the two plots.

The significant result visible on the event-gram is that for the two utterrances, the event-gram is correlated with the average normal listener score, as seen in the blue circles linked by a double arrow. Indeed, for utterance 117, the recognition of consonant /t/ starts to drop, at -2 dB SNR, when the burst above 3 kHz is completely masked by the noise (top right panel of Fig. 2.1). On the event-gram, below -2 dB SNR (blue circle), one can note that the energy of the burst at \( t^* \) decreases, and the burst becomes inaudible (white). A similar relation is seen for utterance 112, but since the energy of the burst is much higher, the /t/ recognition only starts to fall at -15 dB SNR, at which the energy above 3 kHz become sparse and decreases, as seen in the top right panel of Fig. 2.2 and highlighted by the blue circles. A precise and automatic quantification of this correlation for a large numbers of consonants will be future work.

Thus, there is a clear correlation in this example between the variable /t/ confusions and the score for /t/ (step 1, bottom right panel of Figs. 2.1 and 2.2), the strength of the /t/ burst in the AI-gram (step 2, top left panels), the short-time AI value (step 3, bottom left panels), all quantifying the event-gram (step 4, top right panels). This relation generalizes to numerous other /t/ examples and has been here demonstrated for two /tx/ sounds. Because this burst is highly correlated with the human score, it constitutes our model...
of the perceptual cue, denoted event, upon which listeners rely to identify consonant /t/ in noise. The event characterizing the discrimination of /t/ from its competitors is therefore an across-frequency timing coincidence, whereas the sharp onset burst is its physical representation. Finally, the robustness and confusion patterns differences are our proof of the various behavior of utterances of the same CV in noise.

2.2 Truncation Experiment

We may strengthen our conclusions drawn from the correlation of the AI-gram, the confusion patterns and the event-gram. Inspired by the work of Furui [37] and Repp et al. [38], we truncated CV sounds in 5 ms steps, and studied the resulting morphs. The goal was to answer a fundamental research question raised by the four-step analysis of /t/: could the truncation of /t/ cause a morph to /p/, implying that the /t/ event is prefixed to consonant /p/? This conclusion would be in agreement with our observation that some /t/ morph to /p/ when the energy at all frequencies around $t^*$ is masked by the noise.

![Figure 2.3: Change in perception of consonant /t/ due to truncation at 0 dB SNR.](image)

Figure 2.3 for utterance f119ta shows the confusion matrix elements for /t$\alpha$/ pooled over the responses of 10 listeners, as a function of the truncation time from the start of the vowel, at 0 dB SNR in speech-weighted noise. The sound is first heard as /t$\alpha$/ when less than 30 ms of the consonant is truncated. Listeners then reported mostly /p$\alpha$/ for truncation times between 35 and 65 ms, and /b$\alpha$/ for truncations greater than 65 ms. Above this
truncation time, the vowel only may be heard (*). A decrease in the SNR from +12 dB to 0 dB has almost no effect on this conversion (not shown). Every one of the 10 /ta/ tested sounds from our corpus resulted in the same morphs: /ta/ → /pa/ and in most cases further /pa/ → /ba/.

We conclude from the CV-truncation data indicate that the consonant duration is a timing cue used by listeners to distinguish /t/ from /p/. Natural /pa/ utterances morph into /ba/ (not shown), which is consistent with the idea of a hierarchy of speech sounds, clearly present in our /ta/ example. Using such a procedure we independently verified that this burst accounts for the noise robust event corresponding to the discrimination between /t/ and /p/. This is consistent with the results of the AI-gram and event-gram analysis, and will be further verified in the experiment described in this thesis.

2.3 /ma/ - /na/ Discrimination

We have also studied the perceptual difference between /ma/ and /na/. Interestingly, in previous experiments [11], the listeners strongly distinguished the nasal sounds /ma/ and /na/ from all the other consonants which form a phone confusion group. By studying the /ma/ and /na/ utterances of our corpus, we found that the spectrotemporal feature that separates the 2 nasal consonants was likely to be a ≈ 50 ms timing cue between mid and high frequencies. I conducted this pilot study to get insight into this hypothesis. An example of the results is presented in Fig. 2.4.

In the right panel of Fig. 2.4, we see an AI-gram of a /ma/ utterance. Above about 0.6 kHz the energy comes on at the same time. On the left (Fig. 2.4(a)), for /na/, there is a delay in the $F_2$ region relative to the $F_3$ range, due to the convergence of $F_1$ and $F_2$ near 1.2 kHz, as emphasized by the vertical dashed red lines. This rotated V (>), or “boomerang” shaped, formant movement creates a timing distinction between the two frequency
Figure 2.4: Creation of the perceptual timing cue reversing /nA/: (a), (c), and (e), and /mA/: (b), (d), and (f).

ranges, which accounts for the difference in identification of the two sounds.

The middle panels are the confusion patterns control Fig. 2.4(c) and (d), showing the probability of the reported sound with no modification, while the bottom panel Fig. 2.4(e) and (f) gives the score following the removal of the green region. We removed spectrotemporal regions, as shown by the shaded areas, which we have found to be critical for the distinction of those sounds in noise. The shaded (green) region has been removed using the STFT tool to obtain the modified sound, as presented in Chapter. 1 and further detailed in Chapter. 3. The methods for this pilot studies were the same as those for the verification experiment (Chapter. 3), except that the data were not collected in a sound proof booth but in a quiet environment, listeners wearing noise-isolating earphones (Etymotic ER6).

The results demonstrate that, when /nA/ is edited, as shown by the shaded region on the left hand side AI-gram, the sound robustly becomes
/ma/, as reported by our 9 listeners Fig. 2.4(e). The original /na/ was correctly identified as /na/ Fig. 2.4(c) and the modified /na/ utterance, by making all frequencies above 0.6 kHz come on at the same time, produced robust morphing, independent of SNR. When the sound on the right is edited, as defined by the shaded region, /ma/ becomes /na/, as verified by Fig. 2.4(f). At 0 dB SNR /ma/ → /na/ by 8 to 1 (bottom left shows \( \approx 90\% \) versus \( \approx 10\% \) at -2 dB SNR). However, at +10 dB SNR, the modified /ma/ recovers and is heard 60% of the time relative to a 40% /na/ recognition. As seen on the top right AI-gram Fig. 2.4(b), this inversion is due to an unmasked burst of speech energy (circled in red) around 1.2 kHz, before 32 cs and preceding the shaded patch. Casual listening experiments confirm that when the shaded region is fully masked to the left by 10-20 ms, the conversion was complete at all SNRs. By masking this specific part of the /ma/ sound, we created an artificial timing cue, between high frequencies, the onset of \( F_1 \) and \( F_2 \), and the first two formants joining frequency, that simulated the one present in a natural /na/.

All preliminary results, presented in Figs. 2.5-2.7, show that all the /na/ utterances morphed to /ma/, whereas some /ma/ did not morph to /na/. The control confusion patterns are always presented above those for the modified file (right-hand side) and its corresponding AI-gram (left-hand side).

The difference between the regions removed for morphing /ma/ utterances, such as f105ma and f101ma, versus utterances that did not morph to /na/, is the high cut-off frequency of the region. For nonmorphing utterances, such as f103ma, m102ma, m111ma, and m112ma, the upper cut-off was located above the \( F_2 \) onset at 0 dB SNR (not shown for all utterances), suppressing any possible timing distinction between this onset and the joining of \( F_1 \) and \( F_2 \). It also suppressed the possibility of having created a formant transition. Indeed, for morphing utterances, it is also likely that removing the region created an artificial formant transition, due to a new
F2 onset frequency. This hypothesis will be further discussed in Chapter 5.

This midfrequency timing event, or formant transitions could explain the distinction between /ma/-/na/ in noise, and, in part, the insensitivity to spectral reduction schemes, such as Shannon et al. [39] and Loizou et al. [40]. Indeed, spectral reductions smooth over the speech spectrum and remove detailed spectral landmarks, leaving the timing information largely intact and detectable.

We concluded that the acoustic feature leading to the perceptual difference between /ma/-/na/ could be a cross-frequency timing cue, present in the case of /na/, but not in /ma/ utterances. Therefore, part of the verification experiment aimed at collecting more data to confirm this hypothesis. However, as presented in Chapter 4, there seem to be additional robust-to-noise cues to be taken into account for this discrimination, such as the nasal murmur, formants transitions, and/or formants onset frequencies.
Figure 2.6: /ma/ utterances unaffected by the region removal, likely due to a too high upper cut-off.
Figure 2.7: /na/ utterances, all morphing to /ma/. 
To summarize, the overall approach taken in the preliminary studies aims at directly relating the AI-gram, as a model of speech audibility in noise, to the confusion discrimination measure, the confusion patterns versus SNR. This approach is novel, and represents my main contribution toward solving the speech robustness problem. It successfully led to the identification of the /t/ event, and gave insight into the discrimination between /m/ and /n/.

Therefore, the verification experiment, described in the following chapters, naturally follows this analysis. It investigates the responses of CV syllables from many talkers that have been modified using the STFT processing (as described in Chapter 3) to demonstrate further the impact of modifying the acoustic correlates of events.
CHAPTER 3

METHODS

3.1 Overview

This chapter describes the methods of the verification experiment. This experiment verifies the physical support of some perceptual features previously identified and reports on the different morphs (when a sound is reported as another) associated with the modified files. It also reports on the regions having no perceptual consequences. The following sections outline the methods used, with special focus on the choice of the utterances used. The results will then be compared to the four-step analysis method in the case of consonant /t/. Based on the analysis described in Chapter 2, the verification experiment aimed at analyzing the perceptual responses of the following:

- /t/, /p/, and /k/ utterances where the high-frequency burst has been removed. The pilot study showed a high correlation between the /t/ burst and its recognition, therefore the verification was extended to the other two unvoiced stop consonants.

- Modified /mA/ - /nA/, /mA/ - /nA/ utterances, to explain the discrimination between the two nasals. For /m/ utterances, a delay was introduced in the F1-F2 (for vowel /α/) or F2-F3 (for vowel /ι/) frequency ranges. For /n/ utterances, all frequencies above the nasal murmur were edited to come on at the same time for both vowel contexts. The /m/ - /n/ pilot study showed that such processing would lead to a morph in vowel context /α/, therefore a similar removal was applied
to vowel /r/, independently of the potential success of the morph.

- Identify possible forward-masking, or perceptually insensitive regions of energy.

- Identify possible perceptual coarticulation effects for the studied vowels on the consonants included in the experiment.

3.2 My Language Background

The reader must be aware that my first language is French, which could explain possible discrepancies between my optimal modifications and those reported by the subjects. I grew up in the South of France, although I do not have a southern accent. From 1998 to 1999, I lived in New Caledonia before moving to Paris. I have been speaking English for about 10 years now, with more intense practice in the United States for the last 2 years. I also speak Spanish, but I have not practiced recently.

3.3 Subject Selection

Subjects were volunteers from the University of Illinois student and staff population. The participants were required to be between 18 and 23 years of age, have normal hearing (self-reported), and native English speakers. They were compensated $6 per hour for their participation. Ten subjects participated in the experiment, yielding an average of 11.1 presentations per sound per SNR. Table 3.1 summarizes age and number of presentations per subject.

<table>
<thead>
<tr>
<th>Subjects ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>20</td>
<td>22</td>
<td>23</td>
<td>19</td>
<td>20</td>
<td>19</td>
<td>20</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>Presentations</td>
<td>2210</td>
<td>784</td>
<td>1834</td>
<td>2506</td>
<td>1722</td>
<td>2309</td>
<td>1723</td>
<td>1722</td>
<td>2292</td>
<td>1946</td>
</tr>
</tbody>
</table>
3.4 Hardware and Software

The Matlab presentation program, denoted Svan, first written by Andrew Lovitt from our research group, was modified to adapt to the needs of the verification experiment. It was run from a PC (Linux kernel 2.4, Mandrake 9) located outside an acoustic booth (Acoustic Systems model number 27930). Only the keyboard, monitor, headphones, and mouse were inside the booth. Subjects seating in the booth are presented with the speech files through the headphones (Sennheiser HD280 phones), and click on the corresponding file they heard on the user interface (GUI). To prevent any loud sound, the maximum pressure produced was limited to 80 dB sound pressure level (SPL) by an attenuator box located between the soundcard and the headphones. None of the subjects complained about the presentation level, and none asked for any adjustment when suggested.

3.5 Talkers and Utterances

The speech stimuli are taken from the Linguistic Data Consortium (LDC) at the University of Pennsylvania (corpus LDC-2005S22). The corpus is a maximum entropy database of CV, VC, CCV, and VVC utterances. There are 20 talkers with different language backgrounds. The LDC corpus was used in this experiment because it has been used in previous experiments designed by the HSR Group and would facilitate comparisons across experiments results. Five consonants were included in the experiment: /m/, /n/, /p/, /t/, and /k/, followed by either vowel /ɑ/, as in father (and noted “a” when preceded by a talker ID number in this document) or /i/ as in hit (and noted “xi” when preceded by a talker ID number). Talkers’ labels start with either “f” (female talker) or “m” (male talker), followed by the ID number and the CV.

The responses were a closed set of 15 consonants, the same as MN64
experiment but for three consonants: /θ/ (/TH/), /ð/ (/th/), /v/, and ʒ are respectively replaced by /j/, /h/, /f/ , and “vowel only”. Since the morphs are unsure and likely to be listener dependent, it is not possible to have a perfectly unbiased test. Table 3.2, provided by Sandeep Phatak, shows the correspondence between the darpabet, the LDC notation, and the IPA symbols. The consonants and vowels used in the verification experiment are bold.

The following sections describe the utterances selection process.

3.5.1 /m/ and /n/

Nasals are produced by an occlusion of the vocal tract by the lips or the tongue and the opening of the velar port leading to the nasal passages. Concerning articulatory features, all the nasals in English are voiced, the place of articulation of /m/ is bilabial whereas /n/ is alveolar. When the vocal tract is closed and the nasal passage open, the air radiating from the nose is called the nasal murmur. Since the shape of the pharyngeal - nasal passage does not change with the sound the murmur remain almost the same for all nasals. This murmur is stronger in amplitude below 500 Hz. Therefore, even if the length of the vocal tract used as a secondary branch varies and is longer for /m/ than for /n/, there does not seem to be any robust-to-noise difference in the murmur spectrum. The transition from the consonant to the vowel needs to be investigated to account for the differences between these two nasals. Unlike stop consonants production, /m/ and /n/ nasalize the following vowel. Therefore, we expect to see perceptual differences with respect to the vowel. Therefore, two vowels were included in the experiment, /a/ and /i/, for their difference in duration and formants locations. The typical formant values, taken from [41] are shown in Table 3.3. One can note the significant difference for F1 and F2 values, whereas F3 is the same for both vowels.
Table 3.2: This table provides the IPA symbol for each of the 1 and 2 character strings used to generate the names of the LDC sound files.

<table>
<thead>
<tr>
<th>Example</th>
<th>Darpabet</th>
<th>LDCbet</th>
<th>IPAbet</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/ as in hat</td>
<td>æ</td>
<td>xq</td>
<td>æ</td>
</tr>
<tr>
<td>/u/ as in hut</td>
<td>A</td>
<td>xa</td>
<td>Æ</td>
</tr>
<tr>
<td>/ch/ as in church</td>
<td>C</td>
<td>xc</td>
<td>/v/c / ts</td>
</tr>
<tr>
<td>/th/ as in this</td>
<td>D</td>
<td>xd</td>
<td>ð</td>
</tr>
<tr>
<td>/e/ as in bet</td>
<td>E</td>
<td>xe</td>
<td>ð</td>
</tr>
<tr>
<td>/eg/ as in sing</td>
<td>G</td>
<td>xg</td>
<td>ð</td>
</tr>
<tr>
<td>/t/ as in hit</td>
<td>I</td>
<td>xi</td>
<td>ı</td>
</tr>
<tr>
<td>/j/ as in judge</td>
<td>J</td>
<td>xj</td>
<td>/v/ j / d5</td>
</tr>
<tr>
<td>/ow/ as in boy</td>
<td>O</td>
<td>xo</td>
<td>ò</td>
</tr>
<tr>
<td>/sh/ as in she</td>
<td>S</td>
<td>xs</td>
<td>/s/ / j</td>
</tr>
<tr>
<td>/th/ as in think</td>
<td>T</td>
<td>xt</td>
<td>ð</td>
</tr>
<tr>
<td>/u/ as in put</td>
<td>Ú</td>
<td>xu</td>
<td>ð</td>
</tr>
<tr>
<td>/ow/ as in how</td>
<td>W</td>
<td>xw</td>
<td>ao</td>
</tr>
<tr>
<td>/y/ as in why</td>
<td>Y</td>
<td>xy</td>
<td>aı</td>
</tr>
<tr>
<td>/s/ as in pleasure</td>
<td>Z</td>
<td>xz</td>
<td>/v/z / z</td>
</tr>
<tr>
<td>/a/ as in cot</td>
<td>a</td>
<td>a</td>
<td>a / a:</td>
</tr>
<tr>
<td>/b/ as in bee</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>/au/ as in caught</td>
<td>c</td>
<td>c</td>
<td>ð / æ</td>
</tr>
<tr>
<td>/d/ as in dog</td>
<td>d</td>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>/ai/ as in bait</td>
<td>e</td>
<td>e</td>
<td>æı</td>
</tr>
<tr>
<td>/f/ as in fish</td>
<td>f</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>/g/ as in dog</td>
<td>g</td>
<td>g</td>
<td>g</td>
</tr>
<tr>
<td>/h/ as in he</td>
<td>h</td>
<td>h</td>
<td>h</td>
</tr>
<tr>
<td>/ee/ as in bee</td>
<td>i</td>
<td>i</td>
<td>i / i:</td>
</tr>
<tr>
<td>/c/ as in cat</td>
<td>k</td>
<td>k</td>
<td>k</td>
</tr>
<tr>
<td>/l/ as in look</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>/m/ as in man</td>
<td>m</td>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>/n/ as in man</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>/oa/ as in boat</td>
<td>o</td>
<td>o</td>
<td>o / o:0</td>
</tr>
<tr>
<td>/p/ as in pen</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>/t/ as in real</td>
<td>r</td>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>/s/ as in see</td>
<td>s</td>
<td>s</td>
<td>s</td>
</tr>
<tr>
<td>/t/ as in cat</td>
<td>t</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>/oo/ as in boo</td>
<td>u</td>
<td>u</td>
<td>u / u:</td>
</tr>
<tr>
<td>/y/ as in vow</td>
<td>v</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>/w/ as in win</td>
<td>w</td>
<td>w</td>
<td>w</td>
</tr>
<tr>
<td>/y/ as in you</td>
<td>y</td>
<td>y</td>
<td>y / j</td>
</tr>
<tr>
<td>/z/ as in zoo</td>
<td>z</td>
<td>z</td>
<td>z</td>
</tr>
</tbody>
</table>
Table 3.3: Typical vowel formant values.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>/a/</th>
<th>/i/</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>750</td>
<td>400</td>
</tr>
<tr>
<td>F2</td>
<td>1100</td>
<td>1900</td>
</tr>
<tr>
<td>F3</td>
<td>2600</td>
<td>2600</td>
</tr>
</tbody>
</table>

For each utterance starting with /m/ or /n/, there are five modifications: the reference “hand-picked” modification, corresponding to my hearing of the morph following the modifications. The reference is therefore based on what I heard and may not correspond to what English speaking listeners would hear in all cases, hence, the inclusion of other timing delays in the experiment. The four additional modifications changed the duration of the region according to Table 3.4. In each case, the frequency cutoff of the removed region was the same, only the right boundary in time was increased or decreased by 10 and 20 ms.

Table 3.4: Example labeling of talker ID numbers for /m/-/n/ files.

<table>
<thead>
<tr>
<th>Timing</th>
<th>Original</th>
<th>-20 ms</th>
<th>-10 ms</th>
<th>Ref</th>
<th>+10 ms</th>
<th>+20 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talker ID</td>
<td>f101</td>
<td>f301</td>
<td>f201</td>
<td>f401</td>
<td>f501</td>
<td>f601</td>
</tr>
</tbody>
</table>

3.5.2 /p/, /t/, and /k/

Table 3.5 summarizes global information from previous experiments run by the HSR Group: UIUC-S04, MN05 (repeat of Miller and Nicely 1955 experiment), and the truncation experiment.

One can observe that there are discrepancies between the morph in the truncation experiment and S04, as for talker f105, for example. Indeed, for S04, the /t/ burst is not always masked by the noise and constitutes key information even at very low SNRs. In the truncation experiment, the burst is truncated more and more, until it disappears and gives rise to a morph. Therefore, editing the sound by removing the burst should lead to a result consistent with the truncation experiment results.
### Table 3.5: Confusions for /t/ and /p/ utterances common in various experiments.

<table>
<thead>
<tr>
<th></th>
<th>T. COT</th>
<th>M.T</th>
<th>Competitor MN05</th>
<th>M.MN05</th>
<th>Competitor S04</th>
<th>M.S04</th>
</tr>
</thead>
<tbody>
<tr>
<td>f101pa</td>
<td>70 ms</td>
<td>/b/</td>
<td>/b/&gt;/k/</td>
<td>-</td>
<td>abs</td>
<td>abs</td>
</tr>
<tr>
<td>f103pa</td>
<td>≈ 60 ms</td>
<td>VO</td>
<td>/k/&gt;/t/&gt;/f/</td>
<td>/k/</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f106pa</td>
<td>≈ 65 ms</td>
<td>/b/</td>
<td>/k/&gt;/f/</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f109pa</td>
<td>≈ 25 ms</td>
<td>/b/</td>
<td>/f/&gt;/b/, /k/, /t/</td>
<td>/f/</td>
<td>/b/, /f/, /v/</td>
<td>/b/&gt;/v/</td>
</tr>
<tr>
<td>f113pa</td>
<td>from 85 ms to 105 ms</td>
<td>/m/</td>
<td>/k/, /m/, /n/</td>
<td>/m/, /n/</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m104pa</td>
<td>30 ms</td>
<td>VO</td>
<td>/f/, /b/, /n/</td>
<td>-</td>
<td>/v/, /b/, /t/</td>
<td>/b/, /t/</td>
</tr>
<tr>
<td>m107pa</td>
<td>65 ms +</td>
<td>VO</td>
<td>/k/&gt;/t/</td>
<td>/k/&gt;/t/</td>
<td>/k/&gt;/f/, /m/, /b/</td>
<td>&gt; /m/, /b/</td>
</tr>
<tr>
<td>m112pa</td>
<td>50 ms +</td>
<td>VO</td>
<td>/f/, /k/, /b/</td>
<td>-</td>
<td>/b/&gt;/f&gt;/k/</td>
<td>/b/&gt;/f/</td>
</tr>
<tr>
<td>m115pa</td>
<td>60 ms</td>
<td>VO</td>
<td>/k/, /t/</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m117pa</td>
<td>50 ms +</td>
<td>VO</td>
<td>/k/, /t&gt;/m/</td>
<td>/t/</td>
<td>abs</td>
<td>abs</td>
</tr>
<tr>
<td>f105ta</td>
<td>25 ms</td>
<td>/pa/</td>
<td>/p/, /k/</td>
<td>/p/, /k/</td>
<td>/d/</td>
<td>-</td>
</tr>
<tr>
<td>f109ta</td>
<td>20 ms</td>
<td>/pa/</td>
<td>/p/, /k&gt;/f/</td>
<td>/p/, /k/</td>
<td>/d/&gt; /g/, /k/</td>
<td>-</td>
</tr>
<tr>
<td>f113ta</td>
<td>35 ms</td>
<td>/pa/</td>
<td>/p/, /k/</td>
<td>/p/, /k/</td>
<td>/z/, /d/</td>
<td>-</td>
</tr>
<tr>
<td>f119ta</td>
<td>30 ms</td>
<td>/pa/ to /ba/</td>
<td>/k/</td>
<td>/k/</td>
<td>/d/, /TH/</td>
<td>-</td>
</tr>
<tr>
<td>m102ta</td>
<td>10-15 ms</td>
<td>/pa/ /ma/)</td>
<td>/p/</td>
<td>/k/</td>
<td>-</td>
<td>abs</td>
</tr>
<tr>
<td>m104ta</td>
<td>20 ms</td>
<td>/pa/ VO</td>
<td>/p/&gt;/k/</td>
<td>/p/</td>
<td>/k/, /p/, /d/&gt;/g/</td>
<td>/k/, /p/, /d/</td>
</tr>
<tr>
<td>m107ta</td>
<td>15 ms</td>
<td>/pa/</td>
<td>/p/&gt;/k/</td>
<td>/p/&gt;/k/</td>
<td>abs</td>
<td>abs</td>
</tr>
<tr>
<td>m111ta</td>
<td>35 ms</td>
<td>/pa/</td>
<td>/p/&gt;/k/, /m/</td>
<td>/p/&gt;/k/, /m/</td>
<td>/k/</td>
<td>-</td>
</tr>
<tr>
<td>m117ta</td>
<td>10-15 ms</td>
<td>/pa/</td>
<td>/k/&gt;/p/</td>
<td>/k/</td>
<td>abs</td>
<td>abs</td>
</tr>
<tr>
<td>m120ta</td>
<td>45 ms +</td>
<td>/pa/</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
For vowel /\(\alpha/\)

The strategy is to choose utterances that were present in the three experiments, pronounced by three males and three females, resulting in at least six utterances for each CV, in order to collect maximum information for those utterances and compare the results for all experiments. The priority is given to sounds that have morphs in either or both MN05 and S04, and one sound having weak confusions if it exists. To some extent, a variety of main competitors was also a criterion for the selection. The utterances selected appear in bold in Table 3.5. The truncation crossover time (T.COT) corresponds to the time from the beginning of the consonant at which the competitor’s score overcomes that of the original consonant. This values were used as guidelines to establish durations for the removed regions in the verification experiment. The morph in the truncation experiment (denoted M.T) is the sound that most listeners report when the target sound’s score start to decrease. M.MN05 is the morph in MN05, if any. Similar for M.S04. Each entry is this column therefore corresponds to competitors consonants, sometimes ordered by the “>” sign depending on their highest score. A dash means that there is no entry for that column, i.e., the sound does not have strong competitors or does not morph. “Abs” means that the sound is not present in the specified experiment.

For vowel /\(\iota/\)

For CVs that were not in the truncation experiment, for example when starting with /k/ or containing vowel /\(\iota/\), utterances present in both UIUC-S04 and MN05 were first selected. There were 19 utterances of each CV in MN05, and 15 were common with UIUC-S04. Additional utterances were selected because they presented interesting and various properties, such as various competitors, or strong robustness.
3.5.3 List of utterances

The final list of utterances selected is presented in Table 3.6. The files previously identified are bold.

Table 3.6: Utterances selected for the verification experiment.

<table>
<thead>
<tr>
<th>Talker</th>
<th>ta</th>
<th>ka</th>
<th>pa</th>
<th>ma</th>
<th>na</th>
<th>ti</th>
<th>ki</th>
<th>pi</th>
<th>mi</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>f101</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f103</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>f105</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>f106</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f108</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f109</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f113</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>f119</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>m102</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>m104</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m107</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m111</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m112</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m114</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>m115</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m117</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m118</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m120</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

3.6 Noise and SNR

The masking noise used was speech-weighted noise, the same used in UIUC-S04 [2] and described in Fig. 1.2. Six SNRs conditions were tested: [-22, -20, -16, -10, -2, 10] dB SNR. These values are the same as in UIUC-S04, except for the quiet condition, replaced by 10 dB to limit any possible artifacts in quiet condition for the modified files. The control conditions (unmodified files) were presented at the same SNRs. Quiet condition scores, drawn from UIUC-S04 results, were all equal to 1, except for m118pxi: 94%, f113txi: 88% (confused with /fa/ and /θa/ 5% each), and f119txi: 94% (confused with /fa/ 6%). The noise generation, noise filtering, and SNR calculation
procedures were all the same as in UIUC-S04 and are described in [2].

3.7 Modification of the Speech Files

The modification of the speech files was done using the Matlab program Beren, version beren.0.9.8d, which enables the user to select a spectrotemporal region of the speech on an AI-gram, and scale this region up or down in the STFT domain, described in Section 1.7. A screen snapshot of Beren with an original file, and the corresponding modified version, is shown in Fig. 3.1. More examples of resulting AI-grams are displayed in Chapter 4. All selected regions in the verification experiment were scaled to 0 (i.e., the selected region was always removed). The processing is done on the clean speech, and noise is then added to the modified file, the SNR calculation being based on the original file. The modifications were hand-picked and created to study the effect of different regions removal, according to the strategy stated in Section 3.1. Each file present in the experiment was listened to by me at several SNRs, then saved at 30 dB SNR, equivalent to quiet condition. The AI-gram of the modified speech sound could be displayed on the right-hand side of the screen to verify that no energy was left in the selected region.

3.8 Experiment Structure

This section describes the structure of the verification experiment. It was administered in 1- to 2-h sessions at a time. To avoid fatigue and boredom effects, no subject was tested for more than 2 h per day.

Each new subject had to go first through the practice experiment. The training GUI was the same as the experimental GUI to familiarize the subject with the CVs used in the test. All possible responses speech sounds were included in the practice session, so that the subject would be able to
Figure 3.1: Example of a /na/ modification using Beren. The left-hand side displays an AI-gram of f105na a -2 dB SNR, and the right-hand side an AI-gram of the modified file such that the shaded (green) region has been removed. The curve just below the AI-gram is the instantaneous AI, while the bottom curve displays the confusion patterns for this utterance as measured by experiment UIUC-S04.
correctly make the distinction between all possible CVs displayed on the screen. The speech sounds were only presented in quiet condition. It was recommended to take two sessions of the practice, each containing 122 presentations. The subject could take more practice sessions depending on their score, and how comfortable they felt with the identification of the sounds. This training session provided feedback to the user by highlighting the correct answer in red after each sound selection. The listeners were free to repeat a sound at will using a “Repeat” button. Words were displayed below each consonant or vowel (except for vowel only case) to help the subjects identify the consonants and vowels. The subject could also see how many sounds had been heard in the current session so that they knew the number of trials that remained.

Following the practice sessions, the subject was allowed to proceed with the experiment, where both original and modified speech sounds were mixed with masking noise. This portion of this experiment did not provide feedback. The presentations were randomized, and 122 sounds at various SNRs were presented per session. The total number of presentations was 19048, corresponding to an average of 11.1 presentations per sound at each SNR, averaged over all subjects, since not all subjects completed the experiment. It would take an average subject about 5 h, practice included, to complete the experiment. The randomized order of file presentation was different for every subject and explains why some subjects listened to some sounds several times. The GUI display is shown in Fig. 3.2. A “No Idea” button could be hit when subjects did not hear any speech or could not identify what they heard, even after repeated trials.
Figure 3.2: User interface (GUI) for the verification experiment and the practice. The consonants are the labeled rows while the columns correspond to the two vowels.
CHAPTER 4

RESULTS

This chapter describes the extent of the STFT modification and the results of the verification experiment. For each figure presented, the top confusion patterns represent the unmodified speech results, as a control. Then each modified sound has the corresponding listeners responses on the left-hand side, and the associated modification on the AI-gram displayed on the right-hand side. The fractions displayed at the different tested SNRs represent the number of times the sound was heard over the total number of presentations at this SNR. The AI-grams are displayed for the original file, and for each modification, the shaded (green) region was removed. Therefore, we can see what the spectrotemporal distribution was in the region, and we can mentally remove the shaded (green) region to obtain an AI-gram of the modified CV. Figure 4.1 displays the color code used in the confusion patterns plots, and used throughout this document.

![Figure 4.1: Legend used in the confusion patterns plots.](image-url)
The main results can be summarized as follows:

- /t/, /k/ utterances where highly sensitive to the removal of the high-frequency burst, leading to /p/ or /h/ morphs in most cases.
- /p/ utterances were largely insensitive to such time-frequency editing.
- /ma/ very rarely morphed to /na/ by introducing a delay in the $F_1$-$F_2$ range, whereas /na/ strongly morphed to /ma/ when all frequencies were edited to have their onset aligned.
- Conversely, /mi/ morphed to /ni/ utterances when a delay in the $F_2$-$F_3$ range was introduced, whereas modified /mi/ utterances did not morph to /mi/.
- Some regions after the /p/ or /k/ burst onset had no effect on recognition. This is consistent with the hypothesis of forward-masking, but could also be due to the irrelevance of the selected region for the consonant identification.
- Strong perceptual coarticulation effects were found for consonant /k/, /m/ and /n/, leading to different perceptual spectrotemporal regions being significant, and depending on the vowel.

4.1 Removing a Region

As described in the Chapter 1, the STFT modification method enables us to remove a spectrotemporal region selected on the AI-gram so that the region becomes inaudible. Example AI-grams of the modified sounds are given in Fig. 4.2 - 4.6 for each consonant in the verification experiment at 0, 10, and 25 dB SNR. For each block of four panels, we can see an AI-gram of the original utterance, where the green region will be edited out to obtain the sound whose AI-grams are displayed on its right, along with
Figure 4.2: Modifications for f108txi.

Figure 4.3: Modifications for f109mxi.

Figure 4.4: Modifications for m102na.

Figure 4.5: Modifications for m111ka.

Figure 4.6: Modifications for m118pa.
their instantaneous AI below. We can see that at even high SNRs, the modification does erase the speech audible completely, due to the use of Kaiser window (the length of the window was 256, and $\beta = 10$ to obtain a $-80$ dB sidebands attenuation). There is slight delay between low and high frequencies at 25 dB as in the very right panel of Fig. 4.3. This is not due to the STFT modification but the filters used to compute the AI-gram, and should not have any audible effect. The main conclusion is that indeed when a region was selected in the experiment, there is no doubt that it was inaudible for the listeners. However, I chose not to test the subjects above 10 dB SNR, to avoid any artifact due to abrupt onset of energy, which could have created perceptual cues that should not be audible, biasing the results.

### 4.2 Summary of /t/ Results

When the /t/ event is removed, for the majority of the files included in the experiment, either /p/ or /h/ were reported. Consonant /t/ is rarely reported when the burst is totally removed. Utterances f109ta, f113ta, f113txi, and m111ta presented a strong /p/ morph (Appendix A, pp. 103, 106, and 121). The burst for these utterances is short duration (30 ms or less), followed by high frequency energy. The first three utterances present a decrease in the /p/ recognition when the energy following the burst is removed. An example of results is given in Fig. 4.7 for utterance f113txi. We can hypothesize that humans use this energy has a /p/ cue when the burst is missing. However, /p/ is still identified by some of our listeners, suggesting the existence of a possible secondary cue, such as formants, common for /p/ and /t/. This hypothesis is also supported by f308txi modifications (see Appendix A, p. 101). The top confusion patterns correspond to the control condition (no modification), and below are presented the two modifications on AI-grams, and their responses. On the confusion patterns plot, the number of hits out of the total number of responses is presented at the corresponding SNR.
Figure 4.7: Results for f113txi and its two modified files, morphing to /p/ (middle panels) and then mostly to “vowel only” (bottom panels).

For f109ta (Chapter 5 or Appendix A, p. 103), the /t/ percept is also completely gone and /ta/ morphed to /pa/ in the two cases. In the first modification, f209, only the burst is removed. However, in the second modification f309, both the burst and the energy immediately following, for about 50 ms, are removed and /p/ is still heard. This effect will be further discussed in Chapter 5, and supports the existence of transitional cues.

Utterances m114txi, m118txi, and m120ta, presented strong /p/, /h/, or even “vowel only” (no consonant) confusions (Appendix A, pp. 127, 137, and 140). These utterances did not have any energy following the burst, the duration of which is fairly long compared to that of utterances strongly
morphing to /p/. For m114txi, the duration of the burst was about 50 ms, compared to about 60 ms for m118txi in Fig. 4.8. For this utterance, /t/ is correctly identified when only the first 30 ms of the burst are removed. This shows that the duration of the burst does not play a role for the /t/ identification, rather the abrupt onset of energy is used by our listeners to hear /t/, as shown in the bottom panel of Fig. 4.8. Utterance m120ta presented a long burst of about 80 ms, with no energy in the consonant part after the burst (see Appendix A, p. 140). Other behaviors include that of f119ta, which presents results inconsistent with the truncation data. Indeed the removal of the burst mostly leads to a /k/ morph, whereas /p/ was
expected. This result will be analyzed further in Chapter 5. For utterance f119txi, its modification starts as /p/, then is heard as /b/, most likely due to low frequency energy perceived leading to this voicing confusion (see Appendix A, p. 112).

4.3 Summary of /p/ Results

For the consonant /p/, several modifications were tested: the effect of the burst, located at high versus low frequencies, and potential forward masking. Figure 4.9 shows the modifications of utterance f101pxi and is a typical example of the result and represents how most utterances behaved in the experiment. We can see that removing the burst at either high or all frequencies only slightly impairs the /p/ recognition, as shown in the top two AI-grams and their corresponding confusion patterns. The removal of the energy immediately following the burst created a slight /t/ confusion at 10 dB SNR. Suppressing both the high intensity burst and the energy following it decreased the /p/ identification only at 10 dB SNR, but not significantly. These results are further emphasized in Fig. 4.10, showing again how the /p/ identification was barely affected by the burst removal. Only utterances m107pxi, f109pa and m118pa (see Appendix A, pp. 119, 103, and 135) showed a decrease in recognition when the burst was edited out. This contradictory effect will be further discussed in Chapter 5.

As more energy is removed in the transitional region, as presented in Fig. 4.11, the /p/ recognition decreases until no consonant is heard (grey curve, labeled “*”), as shown in the bottom panel. Most utterances follow the same trend, and /p/ usually starts being confused when energy going into the vowel is removed, as for utterance f105pxi, or f119pa respectively displayed in Appendix A, pp. 98 and 111.
Figure 4.9: Results for f101pxi and its four modified files, either unaffected by the burst removal or insensitive to spectrotemporal regions following it.
Figure 4.10: Results for /p/ unaffected by burst removal.
4.4 Summary of /k/ Results

The modifications done for /kt/ and /ku/ examples separate 1, 2, or 3 different frequency ranges that were assumed to carry relevant perceptual information for the identification of consonant /k/, depending on the high-
est energy regions of the burst. Figure 4.12 summarizes the results for most 
/k/ utterances, when the entire burst was removed. In this case, the con-
trol confusion patterns are given in Appendix A. They all had a high /k/ 
recognition as low as -2 dB SNR at least. We can see that this burst removal 
lead to either a /p/, /h/ morphs, or suppression of the consonant, similar 
to the burst removal for /t/ utterances.

For vowel context /a/, the /k/ burst is divided in two main parts, and 
the perceptual relevance of the two frequency ranges was tested. Two main 
trends come out of these results: for the first group of utterances, the lower 
frequency part, located at about 1.7 kHz, carried most of the /k/ percept. 
For the second group, this effect was less noticeable, and removing this mid-
frequency part did not cause a strong decrease in the /p/ identification. 
Figure 4.13 presents the utterances of the group affected by the lower fre-
quency burst removal, especially for utterances f103ka and f108ka located 
in the middle and right hand side of the figure. Other utterances, such as 
m114ka and m102ka (see Appendix A, pp. 124 and 112) had less energy in 
this lower part, therefore were not as sensitive to its removal.

4.5 Summary of /m/ - /n/ Results

This section describes the results for utterances starting with /m/ or /n/, 
displaying the most compelling examples. Other plots can be found in Ap-
pendix A. A summary of the CVs which morph in this experiment is given 
in Table 4.1.

<table>
<thead>
<tr>
<th>Original CV</th>
<th>/m L</th>
<th>/mw</th>
<th>/ma</th>
<th>/na</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morph</td>
<td>/m</td>
<td>weak /m/ or none</td>
<td>weak /na/ or none</td>
<td>/ma/</td>
</tr>
<tr>
<td>Figure examples</td>
<td>4.17</td>
<td>4.18</td>
<td>4.15, 4.16</td>
<td>4.14</td>
</tr>
</tbody>
</table>
Figure 4.12: Results for some /ki/ utterances leading to a /p/, /h/, or "vowel only" morph.
Figure 4.13: /kə/ utterances affected by low-frequency burst removal.
4.5.1 /ma/ - /na/ results

As shown in Table 4.1, /na/ utterances presented a very strong /ma/ morph, as exemplified in Fig. 4.14. The conditions giving the strongest transformation usually correspond to the regions of longest duration. We can see that the /ma/ morph (red) was very strong, even at high SNRs. The timing/formant information will be further discussed in Chapter 5. The control condition, along with the other timing modifications can be seen in Appendix A. There was initially a natural weak /ma/ morph for utterance m118na (p. 134) and f119na (p. 109), which was strengthened by the modification. Only f113na did not present a strong morph (p. 105) but its /ma/ confusion greatly increased as the region was extended.

Alternatively, most /ma/ utterances in the experiment did not morph to /na/. For utterances such as f105ma, f113ma, or m120ma (Appendix A p. 138), the modification created a slight morph, as shown in Fig. 4.15 for utterances f105ma and f113ma. The top confusion patterns represent the results without modification, /ma/ (red) was identified down to -2 dB SNR. Then, as we move down, for each modification the removed region was extended by 10 ms to the right, as displayed on the AI-grams. We can notice a slight /na/ morph (blue) for modification f505ma, corresponding to an optimal 50-ms delay at -2 dB SNR, as well as for the three last modifications for f113ma. When the region was further extended for utterance f105ma (left-hand side), listeners started reporting /ma/ again, whereas the morph remained very similar for f113ma.

For the other three utterances, f119ma, m102ma, and m118ma, the modification only increased the /n/ confusion, but there was never a strong morph. An example of this effect, with the modification having the lowest /m/ score is displayed in Fig. 4.16, for the widest removed region of utterance m102ma (modification denoted m602ma).
4.5.2 /mI/ - /mI/ results

As shown in Table 4.1, most /mI/ utterances in the experiment presented a strong /mI/ morph. The results were therefore opposite to those of vowel
context /a/, since for /ɪ/, /m/ morphed to /n/, whereas /n/ modifications led to a weak decrease in recognition. Examples of strong morphs for the /mɪ/ utterances are presented in Fig. 4.17. For some utterances, such as m117mxi and 118mxi, the conversion works even for the shortest duration region. For utterance m114mxi, the region seemed to not have been extended enough, and therefore did not lead to a very strong morph. The additional horizontal region around 1 kHz for f105mxi was removed to suppress an
audible artifact that could have altered the results.

Alternatively, /m/ utterances modifications were still identified as /m/ by our listeners, even if the /m/ confusion was increased for f103nxi (Fig. 4.18), m119nxi, m102nxi, m114nxi and m117nxi. The modification for f105nxi had no effect (see p. 97). The complete detailed results are shown in Appendix A.
Figure 4.17: /m/ utterances morphing to /m/.

Figure 4.18: Increase of /m/ confusion for f103nxi.
5.1 Results for /t/

The results presented in this thesis firmly confirm that the high-frequency burst is used by listeners to identify consonant /t/ in noise. There is excellent consistency between the verification and truncation experiments, where in both cases the /t/ recognition significantly decreases or is even absent when the burst is becomes inaudible. This section relates the event-gram analysis and the results from the verification experiment, and further discusses the relation with the truncation data, to explain the /p/ morph as found for most cases.

5.1.1 Four-step analysis assumptions for /t/

The four-step method demonstrates the importance of the /t/ burst for the recognition of /t/. Indeed, when the burst was removed, /t/ was almost never reported. There are many examples where the event-gram demonstrates this correlation with the confusion patterns, as in Fig. 5.1. However, since the event-gram is computed at a particular time $t^*$ (see Chapter 1 for the definition of $t^*$), little is known about the perceptual relevance of other times or about the existence of other cues that also contribute to recognition. All existing four-step analysis plots for /ta/ and /ti/ utterances present in UIUC-S04 are given in Appendix C.
5.1.2 Comparison with truncation data

In general, the results are consistent with the truncation experiment data and may be interpreted as a hierarchy for consonant identification in the perceptual domain, or more precisely, the existence of common features shared by consonants of a confusion group. For example, Fig. 5.2 shows the morph results for the truncation experiment at 0 dB, as a function of truncation time from the beginning of the vowel. The truncated /tʌ/ (green) was heard as /pʌ/ (blue) and sometimes as /kʌ/ (red). This pattern also appears in the verification experiment, as shown in Fig. 5.3. Indeed, for modification f209ta (the middle panel of Fig. 5.3), there is some /k/ confusion, whereas when the energy following the /t/ burst is removed, the /k/ confusion disappears in both experiments.

A similar consistency is also seen for utterances that did not morph to /p/ only, depending on the spectral constitution following the burst. Figure 5.4 displays the results of the truncation experiment, as a function of truncation time. Consonant /t/ (green) mostly morphed to either /p/ (blue) or “vowel only” (black *). The confusion patterns in this case are in agreement with those resulting from the burst removal in the truncation.
Figure 5.2: Truncation of f109ta at 0 dB SNR in speech-weighted noise.

Figure 5.3: Results for f109ta and its two modified files.

experiment, and presented in Fig. 5.5. Indeed, since the energy immediately following the burst is weak, the subjects mostly reported that no consonant was heard (grey).

Next we discuss some minor discrepancies between the truncation data and the results presented in this thesis, in particular for utterance f119ta
Figure 5.4: Truncation of m104ta at 0 dB SNR in speech-weighted noise.

Figure 5.5: Results for m104ta morphing to /h/ and * (vowel only).

as shown in Figs. 5.6 (truncation) and 5.7 (confusion patterns), where the removal of the burst mostly lead to a weaker /p/ morph than expected due to the high /k/ confusion. It is possible that since 70 ms of the burst were removed in the verification experiment, the resulting file is located at at about $t = 0.01$ sec on Fig. 5.6, where the /p/ recognition is already low. Additionally, the cutoff of the region was at about 1.4 kHz in the verification, therefore possible low frequency cues could well play a role, whereas all frequencies were removed in the truncation experiment. Indeed, we can see in Fig. 5.8, which shows an AI-gram at 7 dB SNR, that when the noise is decreased there is some mid/low frequency energy at 0.4-1 kHz which might be biasing our listeners towards /k/ versus /p/.
Figure 5.6: Truncation of f119ta at 0 dB SNR in speech-weighted noise.

Figure 5.7: Results for f119ta and its two modified files, morphing to /p/ and confused with /k/ and /h/.

Figure 5.8: AI-gram of f119ta at 7 dB SNR.

5.1.3 Energy located after the /t/ burst

As reported in Chapter 4 of this thesis, /t/ utterances morphing to /p/ had their /p/ recognition impaired when the energy immediately following the
/t/ burst was removed (as in Fig. 4.7), result consistent with the truncation data previously discussed. The resulting plots are located in Appendix A, pp. 106 and 121. However, f109ta does not follow this trend (Fig. 5.3), since the /p/ recognition did not decrease when this notch of energy was removed. However, when looking at an AI-gram of f109ta at 10 dB SNR in Fig. 5.9, we see a low frequency burst having strong intensity. This cue is very likely what listeners were using primarily in order to identify /p/ at this SNR, prior to the high frequency following the burst. It is worth noticing that this low-frequency burst is not a /t/ cue, since the /t/ recognition is zero at this SNR. This burst is also present at 10 dB SNR for m212txi, justifying the /p/ recognition at this SNR, which decreases quickly because of the nonrobustness of this burst.

In the case of f109ta, the /p/ score barely decreases when the low-frequency burst goes into the noise, again confirming the existence of cues used by listeners to identify /p/ in noise, different than the release burst. Utterances morphing to either /p/ or /h/, which did not present this following energy performed at chance in between /p/, /h/, and “vowel only,” most likely because the additional /p/ cue in f109ta was weak for these utterances. These results are consistent with the /p/ analysis (which follows), showing that the recognition of modified /p/ files in the experiment did not present a morph, and were rarely impaired by the burst removal.
5.2 Results for /p/

As shown in Figs. 4.9 and 4.10, and in many other examples in Appendix A, removing the burst has almost no effect on the /p/ recognition. Other possible robust cues include energy around the onset formant frequencies or formant transitions. Results for f105pxi (see Appendix) suggest that the burst might be a perceptual cue at high SNRs, and that other cues could either be introduced by noise, or be revealed and become more salient due to the masking of other energy regions. Removing the /p/ burst caused a decrease in recognition for utterances f109pa and m107pxi and m118pa (see Appendix A, pp. 103, 119, and 135), as presented in Fig. 5.10. This counter trend is likely to be due to a weaker information in the consonant to vowel transition region, which appears to be a strong perceptual cue for the /p/ sounds in the experiment. Indeed, for utterances f109pa and m107pxi which have a very short duration burst, the event-gram plotted at the time of strongest energy is highly correlated with the human score, as in the case of /t/.

5.3 Results for /k/

For some /ka/ utterances, as presented in Chapter 4, the /k/ recognition mainly depends on the energy of small burst just above the F2 transition. However, a modification is shown in Fig. 5.11, for vowel context /a/ for talker m111, presenting three burst of energy instead of the usual two. We can see that the independent removal any of the three strong energy regions did not impair the /k/ recognition. However, when the whole burst is removed, on the bottom plot, most subjects reported hearing /p/. The simplest interpretation of this is that listeners are integrating the /k/ burst information over frequency, using different frequency ranges if some are missing. The utterance is then heard as /p/ if none of those frequencies were activated.
Figure 5.10: Results for f109pa and m107pxi affected by the burst removal.
Figure 5.11: Results for m111ka and its four modifications.
in time. The importance of the different frequency parts of the /k/ burst could be relevant, depending on how much energy there is in each region.

Another interesting modification is that of f209kxi and f202kxi, presented in Fig. 5.12. The first modification (middle panels) demonstrates that for some /k/ utterances, it is possible to transform the /k/ burst into a /t/ burst, by truncating the lower frequency part of it. This result is consistent with the AI density spectra plot (Fig. 5.13) where we can see that for vowel context /i/, the spectral composition of /k/ burst, at the time of strongest energy, carries some similarities with that of /t/, but spreads lower in frequency. The bottom left panel of Fig. 5.12 is somewhat different from results for other /k/ utterances. Indeed, the /k/ recognition peaks at -10 dB SNR, which implies that the burst is not the only perceptual cue that can be used to identify /k/. As can be seen on the bottom left AI-gram for modification f309kxi, there is still energy in the transitional region which could be responsible for the increase of the /k/ percept at -10 dB SNR. On the other hand, the bottom right panel is consistent with the /p/ and /h/ morphs observed for most utterances when the entire burst is removed, and does not contain any energy in the transition region.

5.4 Discriminations between /p/, /t/, and /k/

Results presented here are consistent with the hypothesis that humans are listening to coincidence in different frequency ranges, at least to discriminate /t/ and /k/. Indeed, when only parts of the high energy locations were removed for /k/ utterances, the consonant would be correctly identified in some cases until all burst information is removed. The same pattern is seen for /t/ utterances, leading to different morphs when the complete burst has been truncated. One possible explanation would be that the brain is looking for coincidence across different channels, possibly acting like a logical state machine, weighing or prioritizing different events in the perceptual space.
Figure 5.12: Results for f109kxi and m102kxi where the /t/ confusion is increased.
Figure 5.13 shows the AI density as a function of frequency at the time $t^*$ at which the burst information is the highest, at 0 dB SNR. This figure shows how the burst energy is distributed over frequency for consonants /k/ and /t/ followed by either /a/ or /i/, since it has been shown to be salient information for identification in noise. We can notice the strong coarticulation effect for /k/, where a first sharp peak appears right about 600 Hz above the average second formant frequency for vowel /a/ (usually around 1100 Hz), spreading for about 1 kHz and centered around 1.7 kHz. It is then followed by two other wider bumps, at about 4.2 kHz to 5.2 kHz, and 6.2 kHz up. For vowel context /i/, the /k/ burst has more energy at high frequencies and does not present this strong decrease between frequency ranges, resembling more the /t/ burst. There is no coarticulation effect for /t/ utterances. The /t/ burst starts at about 2 kHz for both vowel contexts, and the increase is very similar.

Both results are very correlated with Miller and Nicely [11] high-pass (HP, right-hand side) and low-pass (LP, left-hand side) experiments (at 12 dB SNR in white masking noise), shown in Fig. 5.14 for /tA/ and /kA/ average utterances. We can see that in both HP and LP conditions (top panels), /t/ was identified when the high frequencies were audible, corresponding the high-frequency /t/ burst, whereas the /k/ recognition decreases (HP, right) or increases (LP, left) when the peak mentioned above, at about 1.7 kHz, respectively becomes inaudible or audible. Other cues in the formants regions are supported by results for /p/, and some results for /k/ such as m114ka or s309kxi, where /k/ is still identified in some SNR range up to 80% of the time even when the burst has been removed. The unsuccessful decrease of the /p/ recognition for the modified /p/ utterances show that there are cues in the vowel, or transition region, giving a /p/ percept to most listeners, consistent with the bottom panels of Fig. 5.14. We can see that in the HP condition (left-hand side), the recognition greatly decreases
Figure 5.13: AI density at $t^*$ for average /ka/, /ki/, /ta/, and /ti/ utterances.

Figure 5.14: Miller and Nicely 55 filtering data for /t/ (top), /k/ (middle), and /p/ (bottom) measured at 12 dB SNR.
when /pa/ utterances are filtered above 2 kHz, i.e., above $F_2$. We can also notice that the /t/ and /k/ confusions are high when the speech is low-pass filtered (right-hand side), with a cut-off frequency below 600 Hz.

These results are consistent with a possible hierarchy of the consonants, since truncated /t/ sounds morph to /p/. The /p/ vowel/transitional cues are likely to be the primary cue in noise for voiceless stop consonants, and then other cues such as the /t/ or /k/ burst are “added” to this primary cue to differentiate the consonants. Consistent with this view, the vowel cues for consonant /p/ would be considered as the basic perception unit of the /p/-/t/-/k/ confusion group, to which burst cues are added to build the other two voiceless stop consonants. Consonant /p/ could therefore be considered as a “root” consonant, presenting primary cues shared by /t/ and /k/.

### 5.5 Forward Masking

There are many examples tested in the described experiment for which forward masking is a possible explanation, as presented in Fig. 5.15. Contrary to Relkin and Turner [42], Duifhuis, Harris et al., and Delgutte [43–45], collected data supporting the existence of forward masking in the auditory nerve. As stated by Harris et al., forward masking of an auditory fiber’s response is defined by a decrease in the probe response caused by a masking stimuli. The recovery of the probe magnitude is an exponential function of time, and depends on the time interval between probe and masker ($\Delta T$).

They also showed that the recovery time and forward masking magnitude were independent of the masker’s spectrum or absolute level, whereas they were influenced by the discharge rate evoked by the masker [43].

Duifhuis [44] showed that the forward masking function could be modeled by two exponentials: for $\Delta T < 10$ ms, the first component reflects the effect of backward masking and very-short-term effects; for $\Delta T > 10$ ms,
Figure 5.15: Forward masking examples, or possibly removed regions having little relevance for consonant recognition.
the second exponential characterizes the exponential recovery (increment of threshold in dB on a log scale versus $\Delta T$): $e^{-\frac{\Delta T}{\tau}}$, with $\tau$ time constant approximated to 75 ms. This component has been interpreted as an adaptation effect in the literature.

In the verification experiment, removing the energy following the burst (tested for /p/ and /k/) had little perceptual consequence on the identification of the target sound. Such an effect could be due to either forward masking, or to the perceptual irrelevance of the removed region for the identification of the consonant, or a combination of both. Note that for the top three /p/ examples (f401pxi, f303pa, and f306pa), removing both the onset of the burst and the energy following it barely impaired the /p/ recognition; this could just imply that listeners do not use these two cues to identify /p/ in noise. However, for the bottom two examples (m304kxi and 518pxi), other modifications (see Appendix A, pp. 117 and 136) show that the recognition is:

- Slightly impaired when the onset of burst is removed,
- Significantly decreased when the whole burst is removed,
- Unaffected when only the region following the onset of the burst has been edited out.

These results could therefore well be due to a forward-masking effect, with the onset of the burst serving as masker for the energy following it. These results need to be investigated further, e.g., in the case of /t/ utterances, in order to eventually add a forward-masking model to the AI-gram.

### 5.6 Nasals Discrimination

The results described previously show the importance of the vowel following the studied consonant in the /m/-/n/ case. Indeed, similar processing was
done, independently of the vowel, and led to very different results. This section analyzes any possible relation between formants, across-frequency timing cues and energy onset location in the spectrotemporal space with the successful or nonexistent morphs in the described experiment. The results also suggest that the nasal resonance might be a stronger cue for /m\textipa{\textael}/ and /\textipa{\textael}/, whereas /\textipa{n}\textipa{\textael}/ and /\textipa{\textael}/ modifications resulted in strong morphs with the murmur left untouched. This hypothesis could be tested by running the verification on high-pass filtered speech, removing the nasal murmur.

5.6.1 /m\textipa{\textael}/ and /\textipa{n}\textipa{\textael}/

As presented in the results chapter, /\textipa{n}\textipa{\textael}/ utterances strongly morphed to /m\textipa{\textael}/, whereas the modifications for /m\textipa{\textael}/ utterances lead to very weak morphs or increased the /\textipa{n}\textipa{\textael}/ confusion.

<table>
<thead>
<tr>
<th>Utterances</th>
<th>f105na</th>
<th>f113na</th>
<th>f119na</th>
<th>m102na</th>
<th>m118na</th>
<th>m120na</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongest morph</td>
<td>f605ma</td>
<td>f613na</td>
<td>f619na</td>
<td>m602na</td>
<td>m418na</td>
<td>m620na</td>
</tr>
<tr>
<td>Optimal region duration [ms]</td>
<td>70</td>
<td>-</td>
<td>80</td>
<td>60</td>
<td>40</td>
<td>65</td>
</tr>
<tr>
<td>Inversion region duration [ms]</td>
<td>50</td>
<td>-</td>
<td>60</td>
<td>20-30</td>
<td>20</td>
<td>45</td>
</tr>
</tbody>
</table>

Figures 5.16-5.21 display spectrograms of /m\textipa{\textael}/ versus /\textipa{n}\textipa{\textael}/ utterances in the experiment, computed using the program Wavesurfer. From these figures, we see that the transition of the second formants $F_2$ is flat for /m\textipa{\textael}/ utterances and downwards for /\textipa{n}\textipa{\textael}/ utterances. The duration of the transitions vary on a token basis.

Based on the pilot studies described in Chapter 2, the intention of the /\textipa{n}\textipa{\textael}/ modifications was to remove the timing cue created by the joining of $F_1$ and $F_2$ and observe the consequence of such editing. The strongest morph timing information is summarized in Table 5.1, and shows the optimal region timing, which happens to display the same trend as the duration of the $F_2$ formant transition. The “region duration” represents the duration of the removed region giving the strongest morph, referenced with respect to
Figure 5.16: Spectrogram of f105ma and f105na.

Figure 5.17: Spectrogram of f113ma and f113na.

Figure 5.18: Spectrogram of f119ma and f119na.
Figure 5.19: Spectrogram of m102ma and m102na.

Figure 5.20: Spectrogram of m118ma and m118na.

Figure 5.21: Spectrogram of m120ma and m120na.
the onset of the \( F2 \) energy at -2 dB SNR on the AI-gram (which is slightly delayed as compared to a spectrogram display in quiet condition, since some irrelevant information has been masked by noise and is not displayed on the AI-gram).

We can see on the spectrograms that the longer the transition, the longer the region needs to be in the verification experiment in order to obtain a morph. Indeed, when the shaded STFT regions were removed from /na/ utterances, the formant transitions were truncated, becoming nonexistent, and changing the onset frequency of \( F2 \). For example, utterance m118na has a very short transition, and the morph starts when only 20 ms of the transition was removed (See Appendix A), unlike utterance m120na, where the morph was very progressive with respect to the timing of the region, which is consistent with a very slow formant transition. The inversion duration corresponds to the duration of the region at which the /m/ score starts overcoming that of /n/, meaning when the morph starts to occur. This duration mostly happens 20 ms before the optimal timing (highest morph), as shown in Table 5.1. These observations therefore support the hypothesis that listeners use formant transitions to discriminate /ma/ - /na/.

However, listeners could also use a timing cue discrimination, created by across frequency timing differences between the high-frequency and \( F1 - F2 \) ranges, as inferred in [21], [22], and by the preliminary results of Figs. 2.5-2.7. Casual listening experiments show that the morph is not complete if only the \( F1 - F2 \) range (and not the high frequencies) was truncated in /na/ utterances, meaning that formant transitions alone are likely to be insufficient to discriminate /ma/ - /na/. This assumption needs to be further investigated to rule out the sole role of the formant transitions in the /ma/ - /na/ discrimination, and more data needs to be collected.

Consistently with the examples given in the pilot studies presented in Chapter 2, /ma/ rarely morphed to /na/ in the verification experiment.
This observation can rule out the hypothesis that having frequencies coming on at the same time is sufficient to give a /ma/ percept but does not exclude the possibility of across-frequency listening. Indeed, for most /ma/ utterances that did not morph to /na/ and similarly to the pilot studies results, the regions removed presented an upper cut-off frequency higher than the $F2$ onset frequency (as visible on AI-grams at 0 or -2 dB SNR). If the removed region suppresses the $F2$ onset frequency, then only a weak increase in the /na/ confusion was observed; when the onset of $F2$ was not removed, /ma/ utterances would morph to /na/.

This further confirms the importance of $F2$ onset frequency, and its transition, without ruling out the relevance of across frequency timing cues, potentially between high frequencies, aligned with the onset of $F1$ and $F2$, and the artificially created joining time of $F1$ and $F2$.

In my opinion, and given the different results obtained, listeners are likely to be sensitive to the onset frequency of the formants, and compare with the onset timing of other frequency ranges to make a decision. This “onset listening” could be combined with a possible formant tracking and across-frequency timing cues. It is likely that no morph takes place for most /ma/ utterances because of either: the impossibility to increase $F2$ onset frequency in order to create a downward $F2$ transition, resembling that of natural /na/ utterances; or a region cut-off frequency too high, often deleting a potential timing cue marked by the $F2$ onset.

This analysis is also consistent with the previous discussion about the sensitivity to onset transient for consonant /t/. Since the auditory system can distinguish which frequencies came on at specific times, the onset-frequency hypothesis for the /ma/-/na/ discrimination is possible, but it is more likely that we could be sensitive to the microfrequency sweeps that characterize formants transitions and integrate the information over frequency to make the discrimination.
5.6.2 /mI/ and /nI/

The /mI/ to /nI/ morphs are in agreement with the previous /mA/ - /nA/ analysis: however, for this vowel, only /mI/ utterances have an upward going formant transitions, whereas those of /nI/ are flat. Figure 5.22 displays an AI-gram of f105mxi at -10 dB SNR on the left hand side, when the recognition was still 100%. Notice the upward formant transition, and the shaded (green) region containing it to be removed. By comparing this with the AI-gram of the natural /mI/ spoken by the same talker on the right hand side, at the same SNR, one can notice the very flat second formant shape. By removing the shaded block in the natural /mI/, we suppress the formant transition; this modification gives rise to flat formant transitions, and change their onset frequencies, leading to the percept of /nI/.

Moreover, when looking at spectrograms of the original /mI/, the /mI/ and the optimal /mI/ modification (f505mxi) in quiet condition in Fig. 5.23 and Fig. 5.24, we can very clearly see this difference, and how the formants of /mI/ and f505mxi are now shaped the same way. More spectrogram examples are given in Appendix B.

There are several serious shortcomings to auditory formant analysis. Spectrograms give very poor representations of speech in noise, and formants detection in quiet condition tell us nothing about the impact of noise. Figure 5.25 displays a spectrogram of f105mxi at -10 dB SNR, in which we
can verify how difficult it is to draw conclusions. The formant display can be misleading as compared to the AI-gram of Fig. 5.22. The AI-gram is therefore the best tool to reliably use in noise for speech analysis, as the formants computed by Wavesurfer are unreliable, especially at poor SNRs.

To summarize, the /m/-/n/discrimination is vowel dependent. “Across frequency onset listening” coupled with formants transitions seems to be the best explanation of our many results and is likely combined with a relative perceptual weight of the murmur compared to that of the transitional region. Listeners seem sensitive to what frequency channels are turned on at what specific time as compared to other frequency ranges.

The nasal murmur role needs further investigation to be ruled out from the set of robust-to-noise cues discriminating consonants /m/ and /n/.
6.1 Conclusions

The following results of the event verification experiment were reported in this thesis:

- /t/ and /k/ utterances where highly sensitive to the removal of the high-frequency burst, leading to /p/ or /h/ morphs in most cases. The noise-robust /t/ burst has been verified to play a significant role in the /t/ identification. When removed, it lead to mostly /p/, /h/ or no consonant morphs, revealing the presence of common features shared by these consonants.

- /p/ utterances were mostly not sensitive to such editing, which shows that there are cues in the vowel or transition region characterizing /p/. This result supports the possibility of a hierarchy of common cues, where /p/ would be the basic unit, and /t/ or /k/ burst would be added to this common cue to obtain the corresponding consonant.

- /mq/ very rarely morphed to /na/ by introducing a delay in the F1-F2 range, most likely due to the removal of F2 formant transition. On the contrary, /na/ strongly morphed to /mq/ when all frequencies were edited to have their onset aligned, suppressing the downward going F2 transition.

Conversely, /mi/ morphed to /ni/ utterances when a delay in the F2-F3 range was introduced. This processing removed the formants
transitions and made them flat, resembling those of natural /m/ utterances. Modified /m/ utterances did not morph to /m/.

These results could be explained by a combination of effect of formants transitions, onset timing across-frequency cues and murmur.

- Some regions after the /p/ or /k/ burst onset had no effect on recognition, likely due to forward-masking or irrelevance of the selected region for the consonant identification.

- Strong perceptual coarticulation effects were found for consonant /k/, /m/, and /n/, leading to different important perceptual spectro-temporal regions depending on the vowel context.

6.2 Future Work

There are many different aspects of this work that could be continued. I spent most of my research time on the pilot studies that lead to the design of the verification experiment. The experiment could benefit from more talkers and utterances, to rule out any experimental uncertainty. It should also be done in white noise, to compare with MN55 and MN05 experiments. More experimental work needs to be done to obtain more information on the /m/ - /n/ discrimination, such as a repeat of the modification combined with high-pass filtering of the nasal murmur, and where the upper cut-off frequency of the region would be varied. Therefore, iterating the process and adding new conditions to the verification experiment would lead to key new results, additional to those stated in this thesis.

The AI-gram could be improved to give a more accurate representation of the perceptual space. Forward-masking models would be a very interesting addition to the model, as well as a possible detection of air release to further identify the /h/ event. Figure 6.1 shows an example of the limitation of the AI-gram and its representation of the audible speech. Indeed the /p/
modification shows that the removed region has very little energy on the AI-gram at 0 dB, but still had an impact on the /p/ identification, decreasing it at high SNRs, but increasing the /p/ recognition at -10 dB SNR, as shown on the middle left panel. However, when increasing the SNR by 5 dB, as shown on the bottom panel AI-gram, we can now notice the presence of the /p/ burst in the green region, which demonstrates that the /p/ burst was likely to be audible at 0 dB SNR, even if its intensity on the AI-gram is very weak.

Figure 6.1: Results for m107pa, its two modified files and AI-gram at -10 dB SNR.

An extension of the event-gram, such as an event-gram plotted in the region of interest, would be really useful to further analyze the relevance of a spectro-temporal region for the recognition. Automated algorithms
that would find the region of strongest correlation with the human response would probably lead this research to a more complete state, by suppressing the time-consuming activities of eye-picking the regions of interest. This would work only if we assume that the events are energy based, and this would necessitate more features being taken into account other than timing delays, such as for the /m/-/n/ distinction. There is also a need to use contour plots on the event-gram or on the proposed extension in order to quantify more quantitatively the correlation with the human responses. In particular, the 2D gradient of the Event-gram could be a useful measure. When looking at the 3D AI-gram, finding the bottom point of the bowl-shaped AI could also lead to interesting results and could be used to find the region of strongest intensity.

Finding algorithms that would track formants in noise and superimpose them on the AI-gram would also display very useful information, in particular for the identification of potential secondary cues in the formant region.

An experiment of enhancement of leading edges should also be run to verify that enhancing a leading edge can increase the recognition of the boosted consonant, a process that could well be one of the most important applications of this research if used to improve the hearing impaired performance in noise.
APPENDIX A

VERIFICATION RESULTS

This appendix shows the confusion patterns for the original, unmodified speech sounds (top), and for their modifications (middle and bottom panels) (Figs. A.1 - A.66). For each modified CV, the corresponding AI-gram at -2 is displayed, where the shaded region represents the spectrotemporal block that was removed from the original file to create this modified speech sound.
Figure A.1: Results for utterance f101pxi and its modifications.
Figure A.2: Results for utterance f103ka and its modifications.

Figure A.3: Results for utterance f103kxi and its modifications.
Figure A.4: Results for utterance f103nxi and its modifications.
Figure A.5: Results for utterance f103pa and its modifications.

Figure A.6: Results for utterance f103txi and its modifications.
Figure A.7: Results for utterance f105ka and its modifications.
Figure A.8: Results for utterance f105kxi and its modifications.
Figure A.9: Results for utterance f105ma and its modifications.
Figure A.10: Results for utterance f105mxi and its modifications.
Figure A.11: Results for utterance f105na and its modifications.
Figure A.12: Results for utterance f105nxi and its modifications.
Figure A.13: Results for utterance f105pxi and its modifications.

Figure A.14: Results for utterance f105ta and its modifications.
Figure A.15: Results for utterance f105txi and its modifications.

Figure A.16: Results for utterance f106pa and its modifications.
Figure A.17: Results for utterance f108ka and its modifications.

Figure A.18: Results for utterance f108pxi and its modifications.
Figure A.19: Results for utterance f108txi and its modifications.

Figure A.20: Results for utterance f109kxi and its modifications.
Figure A.21: Results for utterance f109mxi and its modifications.
Figure A.22: Results for utterance f109pa and its modifications.

Figure A.23: Results for utterance f109ta and its modifications.
Figure A.24: Results for utterance f113ma and its modifications.
Figure A.25: Results for utterance f113na and its modifications.
Figure A.26: Results for utterance f113ta and its modifications.

Figure A.27: Results for utterance f113txi and its modifications.
Figure A.28: Results for utterance f119ma and its modifications.
Figure A.29: Results for utterance f119mxi and its modifications.
Figure A.30: Results for utterance f119na and its modifications.
Figure A.31: Results for utterance f119nxi and its modifications.
Figure A.32: Results for utterance f119pa and its modifications.

Figure A.33: Results for utterance f119ta and its modifications.
Figure A.34: Results for utterance f119txi and its modifications.

Figure A.35: Results for utterance m102ka and its modifications.
Figure A.36: Results for utterance m102kxi and its modifications.
Figure A.37: Results for utterance m102ma and its modifications.
Figure A.38: Results for utterance m102na and its modifications.
Figure A.39: Results for utterance m102nxi and its modifications.
Figure A.40: Results for utterance m104kxi and its modifications.
Figure A.41: Results for utterance m104ta and its modifications.

Figure A.42: Results for utterance m107pa and its modifications.
Figure A.43: Results for utterance m107pxi and its modifications.
Figure A.44: Results for utterance m111ka and its modifications.
Figure A.45: Results for utterance m111ta and its modifications.
Figure A.46: Results for utterance m112kxi and its modifications.
Figure A.47: Results for utterance m112pa and its modifications.

Figure A.48: Results for utterance m112txi and its modifications.
Figure A.49: Results for utterance m114ka and its modifications.
Figure A.50: Results for utterance m114mxi and its modifications.
Figure A.51: Results for utterance m114nxi and its modifications.
Figure A.52: Results for utterance m114txi and its modifications.

Figure A.53: Results for utterance m115pa and its modifications.
Figure A.54: Results for utterance m115pxi and its modifications.
Figure A.55: Results for utterance m117mxi and its modifications.
Figure A.56: Results for utterance m117nxi and its modifications.
Figure A.57: Results for utterance m117txi and its modifications.
Figure A.58: Results for utterance m118ma and its modifications.
Figure A.59: Results for utterance m118mxi and its modifications.
Figure A.60: Results for utterance m118na and its modifications.
Figure A.61: Results for utterance m118pa and its modifications.
Figure A.62: Results for utterance m118pxi and its modifications.
Figure A.63: Results for utterance m118txi and its modifications.
Figure A.64: Results for utterance m120ma and its modifications.
Figure A.65: Results for utterance m120na and its modifications.
Figure A.66: Results for utterance m120ta and its modifications.
This appendix shows spectrograms for /m/ and /n/ files included in the verification experiment (Figs. B.1 - B.8). Some files were not present in the experiment, and their spectrograms are included as a reference for the other nasal consonant. All spectrograms are therefore for unmodified, original sounds, in quiet condition.
Figure B.1: Spectrogram of f103mxi (not in Verif) and f103nxi.

Figure B.2: Spectrogram of f105mxi and f105nxi.

Figure B.3: Spectrogram of f109mxi and f109nxi (Not in Verif).

Figure B.4: Spectrogram of f119mxi and f119nxi.
Figure B.5: Spectrogram of m102nxi (mxi not in Verif nor UIUC-S04).

Figure B.6: Spectrogram of m114mxi and m114nxi.

Figure B.7: Spectrogram of m117mxi and m117nxi.

Figure B.8: Spectrogram of m118mxi and m118nxi.
APPENDIX C

EVENT-GRAMS

This appendix shows event-grams for /tɪ/ and /ta/ utterances from UIUC-S04 (Figs. C.1 - C.14). All confusion patterns are from UIUC-S04 data, in speech-weighted noise.
Figure C.1: Four-step method for f101txi and f103ta.

Figure C.2: Four-step method for f103txi and f105ta.
Figure C.3: Four-step method for f105txi and f106ta.

Figure C.4: Four-step method for f108ta and f108txi.
Figure C.5: Four-step method for f109ta and f109txi.

Figure C.6: Four-step method for f113ta and f113txi.
Step 2: AI-gram of sf119ta.wav at 0 dB SNR

Step 2: AI-gram of sm111ta.wav at 0 dB SNR

Step 4: Event-gram of sf119ta.wav at $f = 26.5$ cs

Step 4: Event-gram of sm111ta.wav at $f = 27$ cs

Figure C.7: Four-step method for f119ta and f119txi.

Step 3: Integrated AI for sf119txi.wav at 0 dB SNR

Step 3: Integrated AI for sm111ta.wav at 0 dB SNR

Figure C.8: Four-step method for m104ta and m111ta.
Figure C.9: Four-step method for m11txi and m112ta.

Figure C.10: Four-step method for m112txi and m114ta.
Step 2: AI−gram of sm114txi.wav at 0 dB SNR

Step 3: Integrated AI for sm114txi.wav at 0 dB SNR

Step 4: Event−gram of sm114txi.wav at $t^* = 26.75$ cs

Step 1: Confusion patterns for m114tx

Figure C.11: Four-step method for m114txi and m115ta.

Step 3: Integrated AI for sm115ta.wav at 0 dB SNR

Step 4: Event−gram of sm115ta.wav at $t^* = 26.25$ cs

Step 1: Confusion patterns for m115ta

Figure C.12: Four-step method for m115txi and m117txi.
Step 2: AI-gram of sm118ta.wav at 0 dB SNR

Step 3: Integrated AI for sm118ta.wav at 0 dB SNR

Step 4: Event-gram of sm118ta.wav at t

Figure C.13: Four-step method for m118ta and m118txi.

Step 2: AI-gram of sm118txi.wav at 0 dB SNR

Step 3: Integrated AI for sm118txi.wav at 0 dB SNR

Step 4: Event-gram of sm118txi.wav at t

Figure C.14: Four-step method for m120ta and m120txi.
REFERENCES


152


