7

Acoustic Simulations of Cochlear Implants in Human and Machine Hearing Research

Cong-Thanh Do
Idiap Research Institute, Centre du Parc, Martigny
Switzerland

1. Introduction

Cochlear implant is an instrument which can be implanted in the inner ear and can restore partial hearing to profoundly deaf people (Loizou, 1999a) (see Fig. 1). Acoustic simulations of cochlear implants are widely used in cochlear implant research. Basically, they are acoustic signals which simulate what the profoundly deaf people could hear when they wear cochlear implants. Useful conclusions can be deduced from the results of experiments performed with acoustic simulations of cochlear implants. There are two typical applications in cochlear implant research which use acoustic simulations of cochlear implants. In the first one, acoustic simulations of cochlear implants are used to define how many independent channels are needed in order to achieve high levels of speech understanding (Loizou et al., 1999). The second application of acoustic simulations in cochlear implants research is for determining the effect of electrode insertion depth on speech understanding (Baskent & Shannon, 2003; Dorman et al., 1997b). In this chapter, we review briefly these conventional applications of acoustic simulations in cochlear implants research and, on the other hand, introduce novel applications of acoustic simulations of cochlear implants, both in cochlear implants research and in other domains, such as automatic speech recognition (ASR) research. To this end, we present quantitative analyses on the fundamental frequency (F0) of the cochlear implant-like spectrally reduced speech (SRS) which are, essentially, acoustic simulations of cochlear implants (Loizou, 1999a). These analyses support the report of (Zeng et al., 2005), which was based on subjective tests, about the difficulty of cochlear implant users in identifying speakers. Following the results of our analyses, the F0 distortion in state-of-the-art cochlear implants is large when the SRS, which is acoustic simulation of cochlear implant, is synthesized only from subband temporal envelopes (Do, Pastor & Goalic, 2010a). The analyses revealed also a significant reduction of F0 distortion when the frequency modulation is integrated in cochlear implant, as proposed by (Nie et al., 2005). Consequently, the results of such quantitative analyses, performed on relevant acoustic traits, could be exploited to conduct subjective studies in cochlear implant research. On the other hand, we investigate the automatic recognition of the cochlear implant-like SRS. Actually, state-of-the-art ASR systems rely on relevant spectral information, extracted from original speech signals, to recognize input speech in a statistical pattern recognition framework. We show that from certain SRS spectral resolution, it is possible to achieve (automatic) recognition performance as good as that attained with the original clean speech even though the cochlear implant-like SRS is synthesized only from subband temporal envelopes of the original clean speech (Do, Pastor & Goalic, 2010b; Do, Pastor, Le Lan & Goalic, 2010). Basing on this result, a novel framework for noise robust ASR, using cochlear implant-like SRS, has been proposed in (Do et al., 2012). In this novel framework, cochlear implant-like SRS is used in both the training and testing conditions. Experiments show that the (automatic) recognition results are significantly improved, compared to the baseline system which does not employ the SRS (Do et al., 2012).

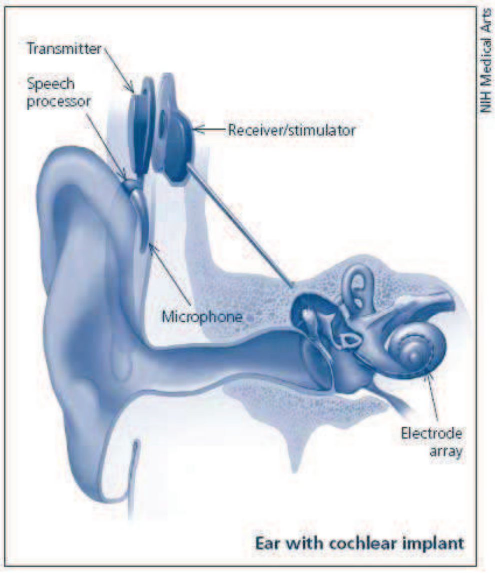


Fig. 1. Cochlear implant is an instrument which can be implanted in the inner ear and can restore partial hearing to profoundly deaf people (Loizou, 1999a) (image source: NIH Medical Arts, USA).

2. Conventional applications of acoustic simulations of cochlear implants

2.1 Researching the number of independent channels needed to achieve high levels of speech understanding

It has been known that the speech understanding does not require highly detailed spectral information of speech signal since much of the information in the speech spectrum is redundant (Dudley, 1939). On the other hand, the fine spectral cues, presented in naturally produced utterances, are not required for speech recognition (Zeng et al., 2005). In general, the signal processing in cochlear implant divides the speech spectrum into several spectral subbands, from 4 to 12 depending on the device, and then transmits the energy in all the subbands. However, knowing how many independent channels are needed in order to achieve high levels of speech understanding is an important issue. In fact, it is difficult to answer this question basing on the results of subjective tests performed by cochlear implant users, since their performance might be affected by many cofounding factors (e.g. number of surviving ganglion cells). For example, if a cochlear implant user obtains poor auditory performance using 4 channels of stimulation, the rationale might be that 4 stimulating channels are not enough but it could probably be blamed for the lack of surviving ganglion cells near the stimulating electrodes (Loizou et al., 1999). Using acoustic simulations can help in separating the rationale coming from the lack of surviving ganglion cells.

In (Shannon et al., 1995), the authors showed that high levels of speech understanding (e.g. 90% correct for sentences) could be achieved using a few as four spectral subbands. The subband temporal envelopes of speech signal were extracted from a small number (1-4) of frequency subbands, and used to modulate noise of the same bandwidth. In their signal processing strategy, the temporal cues within each subband are preserved but the fine spectral cues within each subband are eliminated. In another study, (Dorman et al., 1997a) used subband temporal envelopes to modulate sine waves rather than noise bands, and then, summed these subband modulated signals. As in (Shannon et al., 1995), sentence recognition using four channels was found to be 90% correct.

Typically, in the subjective tests for determining the necessary number of frequency subbands needed for understanding speech, normal hearing listeners listened to stimulus consisting of consonants, vowels, and simple sentences in each of the signal conditions (Faulkner et al., 1997; Loizou et al., 1999; Shannon et al., 2004; 1995). In these tests, consonants and vowels were presented in a random order to each listener. The listeners were instructed to identify the presented stimulus by selecting it from the complete set of vowels or consonants (Shannon et al., 1995). In the sentence recognition tests, sentences were presented once and the listeners were asked to repeat as many words in the sentence as they could. Training (i.e. practice) is a factor that needs to be taken into account when interpreting the speech recognition results mentioned previously, since the normal hearing listeners cannot recognize immediately the speech signals with spectro-temporal distortion. For example, in (Shannon et al., 1995), the listeners were trained on sample conditions to familiarize them with the testing environment; after about two or three sessions, for a total of 8 to 10 hours, when the listeners' performance stabilized, the training can be stopped. No feedback was provided in any of the test conditions.

On the other hand, (Zeng et al., 2005) showed that although subband temporal envelopes (or amplitude modulations - AMs) from a limited number of spectral subbands may be sufficient for speech recognition in quiet, frequency modulations (FMs) significantly enhances speech recognition in noise, as well as speaker and tone recognition. The result of (Zeng et al., 2005) suggested that FM components provide complementary information which supports robust speech recognition under realistic listening situations. A novel signal processing strategy, named FAME (frequency amplitude modulation encoding), was proposed which integrates not only AMs but also FMs into the cochlear implants (Nie et al., 2005; Zeng et al., 2005). More details about the FAME strategy will be presented in section 3.1.

2.2 Simulating the effect of electrodes insertion depth on speech identification accuracy

The second application of acoustic simulations in cochlear implants research is for determining the effect of electrode insertion depth on speech understanding. In cochlear implantation, electrode arrays are inserted only partially to the cochlea, typically 22-30 mm, depending on the state of the cochlea. Acoustic simulations of cochlear implants could be used to simulate the effect of depth of electrode insertion on identification accuracy. In this respect, normal hearing listeners performed identification tasks with an acoustic simulation of cochlear implant whose electrodes are separated by 4 mm (Dorman et al., 1997b). Insertion depth was simulated by outputting sine waves from each channel of the processor at a frequency determined by the cochlear place of electrode inserted 22-25 mm into the cochlea, through the Greenwood's frequency-to-place equation (Greenwood, 1990). The results indicated that simulated insertion depth had a significant effect on speech identification performance (Dorman et al., 1997b).

The mapping of acoustic frequency information onto the appropriate cochlear place is a natural biological function in normal acoustic hearing. However, in cochlear implant, this mapping is controlled by the speech processor. Indeed, the length and insertion depth of the electrode arrays are important factors which determine the cochlear tonotopic range (Baskent & Shannon, 2003). A 25 mm insertion depth of the electrode arrays is usually used in the design of conventional cochlear implant. This design would place the electrodes in a cochlear region corresponding to an acoustic frequency range of 500-6000 Hz. While this mapping preserves the entire range of acoustic frequency information, it also results in a compression of the tonotopic pattern of speech information delivered to the brain. The effects of such a compression of frequency-to-place mapping on speech recognition are studied in (Baskent & Shannon, 2003) using acoustic simulations of cochlear implants.

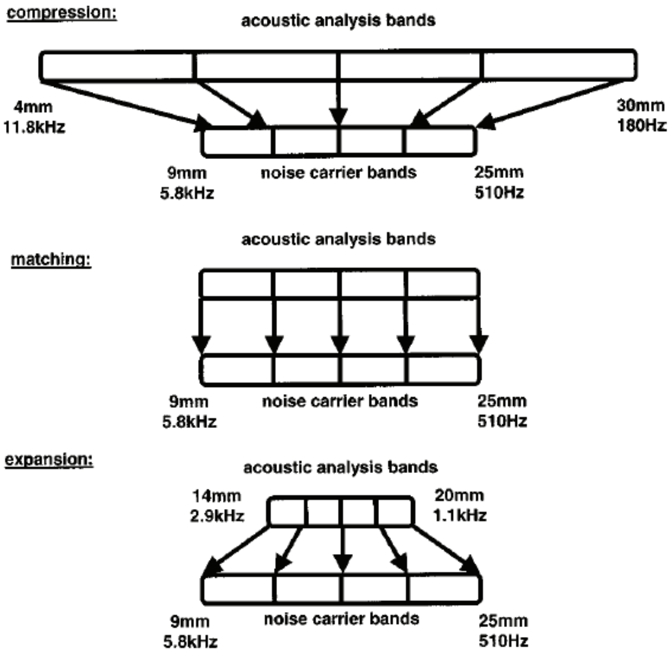


Fig. 2. Frequency-place mapping conditions for 4-channel processor at the simulated 25-mm electrode insertion depth (Baskent & Shannon, 2003). In this condition, the noise carrier bands are fixed (9-25 mm: 510-5800 Hz). The speech envelope was extracted from the analysis subbands and used to modulate the noise carrier subbands. The three panels, in top-down order, show the three mapping conditions, compression, matched and expansion, respectively. In the top panel where there is a compression of +5 mm, the analysis subbands are mapped onto narrower carrier subbands. The middle panel shows the 0 mm condition, in which the analysis and carrier bands are matched. The lower panel shows the -5 mm expansion condition, in which analysis subbands are mapped onto wider carrier subbands.

In (Baskent & Shannon, 2003), speech recognition was measured as a function of linear frequency-place compression and expansion using phoneme and sentence stimulus. Cochlear implant with different number of electrode channels and different electrode insertion depths were simulated by noise-subband acoustic simulations and presented to normal hearing listeners. Indeed, it was found that in the matched condition where a considerable amount of acoustic information was eliminated, speech recognition was generally better than any condition of frequency-place expansion and compression. This result demonstrates the dependency of speech recognition on the mapping of acoustic frequency information onto the appropriate cochlear place (Baskent & Shannon, 2003).

3. Novel applications of acoustic simulations of cochlear implants

In this section, we introduce novel applications of acoustic simulations of cochlear implants, henceforth abbreviated cochlear implant-like spectrally reduced speech (or simply SRS), both in cochlear implant research and other domains, such as automatic speech recognition (ASR). In this respect, section 3.1 presents our SRS synthesis algorithm, based on the frequency amplitude modulation encoding (FAME) algorithm (Nie et al., 2005), that is use through out the rest of this chapter. Section 3.2 introduces an analysis of the speech fundamental frequency (F0) in the SRS and its application in cochlear implant research. In addition, another application of the acoustic simulations of cochlear implants, in ASR, is introduced in section 3.3. Finally, section 4 concludes the chapter.

3.1 Cochlear implant-like SRS synthesis algorithm

A speech signal, $s(t)$, is first decomposed into N subband signals $s_i(t), i = 1, \ldots, N$, by using an analysis filterbank consisting of N bandpass filters with N taking values in {4, 8, 16, 24, 32}. The analysis filterbank is aimed at simulating the motion of the basilar membrane (Kubin & Kleijn, 1999). In this respect, the filterbank consists of nonuniform bandwidth bandpass filters that are linearly spaced on the Bark scale. In the literature, gammatone filters (Patterson et al., 1992) or elliptic filters (Nie et al., 2005; Shannon et al., 1995) have been used to design such a filterbank. In this chapter, each bandpass filter in the filterbank is a second-order elliptic bandpass filter having a minimum stop-band attenuation of 50-dB and a 2-dB peak-to-peak ripple in the pass-band. The lower, upper and central frequencies of the bandpass filters are calculated as in (Gunawan & Ambikairajah, 2004). An example of the analysis filterbank is given in Fig. 3.

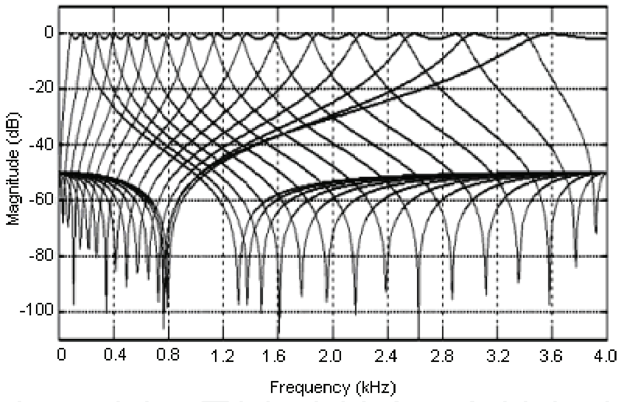


Fig. 3. Frequency response of an analysis filterbank consisting of 16 second-order elliptic bandpass filters used for speech signal decomposition. The speech signal is sampled at 8 kHz.

The subband signals, $s_i(t), i = 1, \ldots, N$, are supposed to follow a model that contains both amplitude and frequency modulations

$$s_i(t) = m_i(t) cos \left(2\pi f_{ci} t + 2\pi \int_0^t g_i(\tau) d\tau + \theta_i \right) \tag{1}$$

where $m_i(t)$ and $g_i(t)$ are the amplitude and frequency modulation components of the i-th subband whereas f_{ci} and θ_i are the i-th subband central frequency and initial phase,

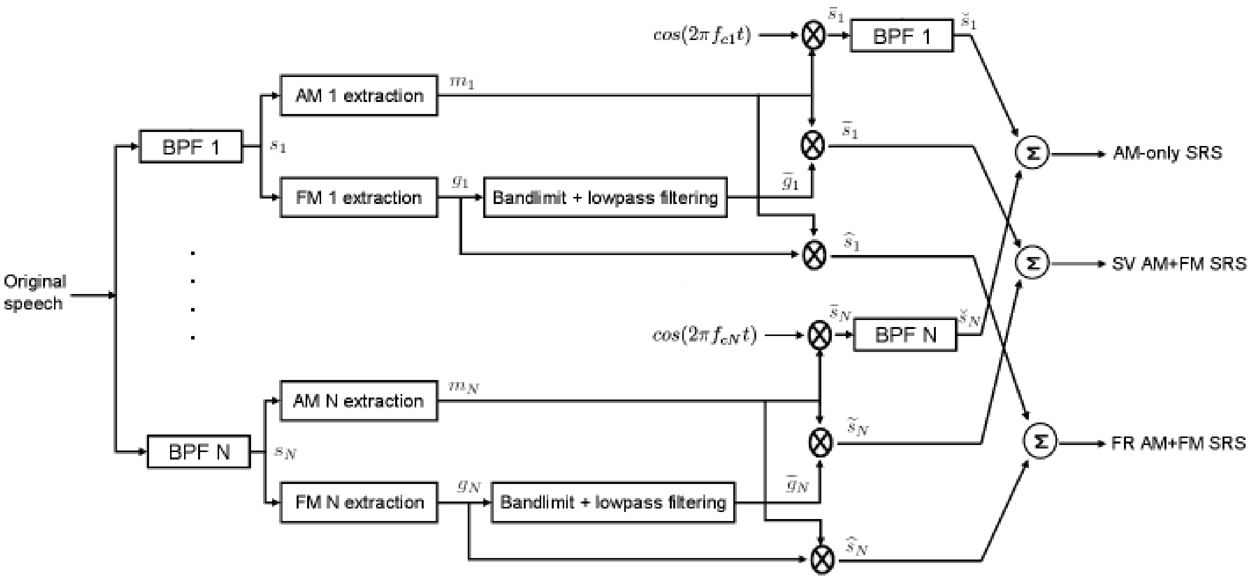


Fig. 4. The SRS synthesis algorithm derived from the Frequency Amplitude Modulation Encoding (FAME) strategy proposed in (Nie et al., 2005). The speech signal is decomposed by a filterbank consisting of N bandpass filters. AM and FM components are extracted from each subband signal. The AM components are used to modulate a fixed sinusoid, or a carrier with or without rapidly varying FM components, then summed up to synthesize the AM-only SRS, the FR AM+FM SRS, or the SV AM+FM SRS, respectively. The $f_{c1}, f_{c2}, \ldots, f_{cN}$ are the central frequencies of the bandpass filters in the analysis filterbank. The FR AM+FM SRS synthesis is our proposition.

respectively. Each AM component m_i of the subband signal s_i is extracted by full-wave rectification and subsequently, lowpass filtering of the subband signal s_i with a 50- or 500-Hz cutoff frequency. In the other way, the subband FM components, g_i, is extracted by first removing the central frequency, f_{ci}, from the subband signal s_i, thanks to a quadrature oscillator consisting of a pair of orthogonal sinusoidal signals whose frequencies equal the central frequency of the i-th subband (Nie et al., 2005). At the outputs of the quadrature oscillator, two lowpass filters are subsequently used to limit the frequency modulation range. The cutoff frequencies of these two lowpass filters equal 500 Hz or the bandwidth of the bandpass filter used in the analysis stage, whenever the latter is less than 500 Hz. The full rate (FR) FM signal, g_i, is then band-limited and filtered by using a 400-Hz cutoff frequency lowpass filter to derive the slowly varying (SV) band-limited FM signal, $\widetilde{g}_i$, as in (Nie et al., 2005). In the synthesis stage, the subband AM signal, m_i, is used to modulate a carrier containing the subband FR FM signal, g_i, or the subband SV FM signal, $\widetilde{g}_i$, to obtain the subband modulated signals, $\widehat{s}_i$, or $\widetilde{s}_i$, respectively. The sum of the subband modulated signals, $\widehat{s}_i, i = 1, \ldots, N$, gives the SRS with FR FM components, called FR AM+FM SRS. Similarly, SV AM+FM SRS is the SRS achieved when summing the subband modulated signals, $\widetilde{s}_i, i = 1, \ldots, N$, which are obtained by modulating the subband SV FMs with the subband AMs. All the lowpass filters used in the AM and FM extractions are fourth-order elliptic lowpass filters having a minimum stop-band attenuation of 50-dB and a 2-dB peak-to-peak ripple in the pass-band.

In the synthesis of the AM-based SRS, the subband modulated signal, $\bar{s}_i$, is obtained by using the subband AM signal, m_i, to modulate a sinusoid whose frequency equals the central frequency, f_{ci}, of the corresponding analysis bandpass filter of the i-th subband. The subband modulated signal, $\bar{s}_i$, is then spectrally limited by the same bandpass filter used for the original analysis subband to derive the spectrally limited subband modulated signal, $\check{s}_i$ (Shannon et al., 1995). The AM-based SRS is the sum of all the spectrally limited subband modulated signals, $\check{s}_i, i = 1, \ldots, N$. This type of SRS is called the AM-only SRS to indicate that the SRS is synthesized by using only AM cues. This description is summarized by the schema of Fig. 4.

3.2 Normalized mean squared error (NMSE) analysis of F0 in the acoustic simulations of cochlear implants

3.2.1 Motivation for the analysis

It is known that cochlear implant listeners do not have sufficient information about the voice fundamental frequency (F0) for their speech recognition task. Meanwhile, F0 information is useful for speaker discrimination and provides critical cues for the processing of prosodic information. Most present-day cochlear implants extract the speech temporal envelope and are not designed to specifically deliver speech F0 information to the listener. Cochlear implant listeners have therefore difficulties to perform the tasks needing F0 information, such as speaker recognition, gender recognition, tone recognition, or intonation recognition, etc (Chatterjee & Peng, 2008; Nie et al., 2005).

The modulation in speech, especially the frequency modulation, is expected to carry speaker-specific information. In (Nie et al., 2005), the authors proposed a speech processing strategy, FAME (for Frequency Amplitude Modulation Encoding), which encodes the speech temporal fine structure by extracting slowly varying band-limited frequency modulations (FMs) and incorporates these components in the cochlear implant. The acoustic simulations of cochlear implant, called spectrally reduced speech (SRS), can be synthesized either by using the FAME strategy or by using conventional AM-based SRS synthesis algorithm (Nie et al., 2005). In (Zeng et al., 2005), the authors performed speaker recognition tests by using these types of SRS. Experimentally, normal hearing listeners achieved significant better speaker recognition scores when listening to the FAME-based SRS, compared to when listening to the SRS synthesized from only AM components. Meanwhile, cochlear implant users could only achieve an average recognition score of 23% when they performed the same speaker recognition tests but listening to the original clean speech (Zeng et al., 2005). These results showed that current cochlear implant users have difficulties in identifying speakers (Zeng et al., 2005).

We thus want to quantitatively clarify the report of (Zeng et al., 2005) on the speaker recognition tasks of cochlear implant users, by performing a comparative study on the speech F0, extracted from the AM-based and FAME-based SRSs. The AM-based and the FAME-based SRSs were synthesized from a set of original clean speech utterances selected from the TI-digits, which is a multi-speaker speech database (Leonard, 1984). The selected speech utterances contain a large number of speakers to take into account the intra-speaker F0 variation. Next, the Normalized Mean Square Errors (NMSEs) between the F0 extracted from the original clean speech and from the SRS were calculated and analyzed.

3.2.2 Data for analysis

A set of 250 utterances was selected from the TI-digits clean speech database. TI-digits is a large speech database of more than 25 thousand connected digits sequences, spoken by over 300 men, women, and children (Leonard, 1984). The data were collected in a quiet environment and digitized at 20 kHz. In this study, the data are downsampled to 8 kHz. These 250 utterances were selected so that they had been spoken by both adults (men, women) and children (boy, girls) speakers. The lengths of the utterances in the set vary from the minimum length (isolated digit sequence) to the maximum length (seven-digit sequence) of the sequences in the TI-digits. The AM-only SRS, the FR AM+FM SRS, and the SV AM+FM SRS were synthesized from these 250 utterances by using the algorithm described in section 3.1. We thus have 250 AM-only SRS, 250 FR AM+FM SRS, and 250 SV AM+FM SRS utterances.

3.2.3 F0 NMSE analysis

The F0 values are extracted from the voiced speech frames of the original clean speech and the SRS utterances by using the Praat software. Praat is a computer program to analyse, manipulate speech signal and compute acoustic features, developed by Boersma and Weenink (Boersma & Weenink, 2009). The Praat F0 extraction algorithm, based on the autocorrelation method, is standard and accurate. A detailed description of the Praat F0 extraction algorithm can be found in (Boersma, 2004). Let $\mathbf{X} = [\mathbf{t}_X \; \mathbf{f}_X]$ and $\mathbf{Y} = [\mathbf{t}_Y \; \mathbf{f}_Y]$ be the vectors extracted from the voiced speech frames (of length 10 ms) of a clean speech utterance and the corresponding SRS utterance, respectively. The vectors $\mathbf{t}_X = [t_X(1), \ldots, t_X(L)]^T$ and $\mathbf{f}_X = [f_X(1), \ldots, f_X(L)]^T$ contain the time instants and the extracted F0 values, respectively, of the voiced speech frames in the original clean speech utterance. Similarly, $\mathbf{t}_Y = [t_Y(1), \ldots, t_Y(M)]^T$ and $\mathbf{f}_Y = [f_Y(1), \ldots, f_Y(M)]^T$ contain the time instants and the extracted F0 values, respectively, of the voiced speech frames in the corresponding SRS utterance. The Praat script for extracting the vectors $\mathbf{X}$ and $\mathbf{Y}$ can be found at `http://www.icp.inpg.fr/~loeven/ScriptsPraat.html`. The superscript ^{T} denotes the transpose whereas L and M are the lengths of the vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively. In general, $\mathbf{X}$ and $\mathbf{Y}$ do not have the same lengths ($L \neq M$) even though the SRS utterance is synthesized from the same original clean speech utterance. The time instant values in $\mathbf{t}_X$ and $\mathbf{t}_Y$ are not identical, either. Without losing the generality, we suppose that $L < M$. In order to correctly calculate the NMSE between the F0 vectors of an original clean speech and a SRS utterance, we calculate the vector $\widehat{\mathbf{Y}} = \left[\widehat{\mathbf{t}}_Y \; \widehat{\mathbf{f}}_Y\right]$ as follows

```
Algorithm 1 Calculating $\widehat{\mathbf{Y}} = \left[\widehat{\mathbf{t}}_Y \; \widehat{\mathbf{f}}_Y\right]$ from $\mathbf{t}_X, \mathbf{t}_Y$ and $\mathbf{f}_Y$
FOR $i = 1 \rightarrow L$
    $\widetilde{j} = \underset{j=1..M}{\arg\min} |t_Y(j) - t_X(i)|$
    $\widehat{t}_Y(i) = t_Y(\widetilde{j})$
    $\widehat{f}_Y(i) = f_Y(\widetilde{j})$
END
```

where $\widehat{\mathbf{t}}_Y = \left[\widehat{t}_Y(1), \ldots, \widehat{t}_Y(L)\right]^T$ and $\widehat{\mathbf{f}}_Y = \left[\widehat{f}_Y(1), \ldots, \widehat{f}_Y(L)\right]^T$ are the new time instants and F0 values vectors of the SRS utterance. The purpose of algorithm 1 is to calculate $\widehat{\mathbf{Y}} = \left[\widehat{\mathbf{t}}_Y \; \widehat{\mathbf{f}}_Y\right]$ from $\mathbf{t}_X, \mathbf{t}_Y$ and $\mathbf{f}_Y$ so that the temporal lags between the identical index elements of $\mathbf{t}_X$ and $\widehat{\mathbf{t}}_Y$ are minimal. The NMSE is then calculated between the two vectors of F0 values, $\mathbf{f}_X$ and $\widehat{\mathbf{f}}_Y$, now having the same length L.

$$\mathrm{NMSE} = 20\log_{10}\left(\frac{1}{L}\sum_{i=1}^{L}\left|\frac{f_X(i)-\widehat{f}_Y(i)}{f_X(i)}\right|^2\right) \tag{2}$$

The temporal lag minimization performed by algorithm 1 makes it possible to achieve an accurate NMSE calculation, following formula (2). The averages of the NMSEs on the selected 250 utterances were calculated and represented as a function of the SRS number of frequency subbands in Fig. 5 and Fig. 6.

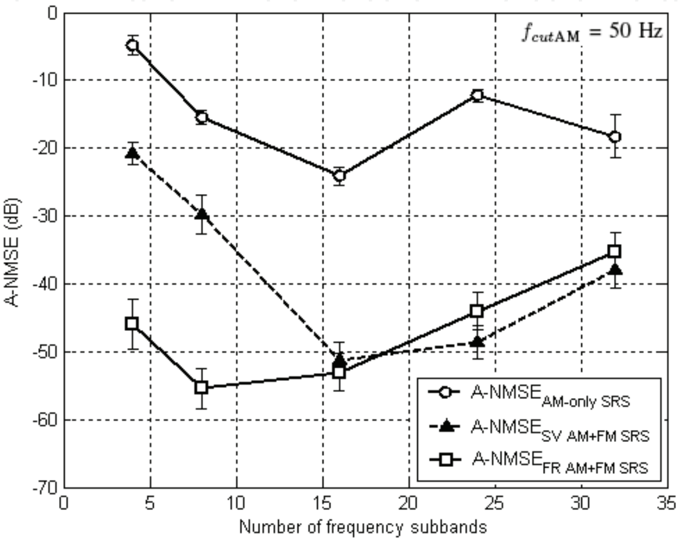


Fig. 5. Averages of the NMSEs (A-NMSEs) between the F0 vectors, extracted from the clean speech utterances and those extracted from the AM-only SRS (A-NMSE$_{\mathrm{AM\text{-}ONLY\ SRS}}$), SV AM+FM SRS (A-NMSE$_{\mathrm{SV\ AM+FM\ SRS}}$), and FR AM+FM SRS (A-NMSE$_{\mathrm{FR\ AM+FM\ SRS}}$), calculated on 250 utterances (section III.A). The AM used for the SRS synthesis were extracted by using a 50-Hz cutoff frequency lowpass filter. Error bars indicate 95% of the confidence interval (Cumming et al., 2007) around each A-NMSE.

The curves in Fig. 5 represent the averages of the NMSEs (A-NMSEs) calculated between the original clean speech F0 vectors and those of the synthesized SRSs, which used 50-Hz cutoff frequency for the AM extraction lowpass filter ($f_{cut\mathrm{AM}}$ = 50 Hz). Similarly, the A-NMSEs calculated between the synthesized SRS, having $f_{cut\mathrm{AM}}$ = 500 Hz, and the original clean speech, are represented by the curves in Fig. 6. Henceforth, we use the terms A-NMSE$_{\mathrm{AM\text{-}ONLY\ SRS}}$, A-NMSE$_{\mathrm{SV\ AM+FM\ SRS}}$, and A-NMSE$_{\mathrm{FR\ AM+FM\ SRS}}$ to designate the A-NMSEs calculated between the F0 vectors of the original clean speech and those of the AM-only SRS, SV AM+FM SRS, and FR AM+FM SRS, respectively. An one-way ANOVA reveals that the overall difference between the A-NMSEs is significant $[F(2,12) = 14.5, p < 0.001]$ for $f_{cut\mathrm{AM}}$ = 50 Hz. Similarly, when $f_{cut\mathrm{AM}}$ = 500 Hz, the A-NMSEs overall difference is also significant $[F(2,12) = 9.8, p < 0.005]$. Further, the A-NMSE$_{\mathrm{AM\text{-}ONLY\ SRS}}$ is significantly greater than the A-NMSE$_{\mathrm{SV\ AM+FM\ SRS}}$ $[F(1,8) = 12.1, p < 0.01]$, and the A-NMSE$_{\mathrm{FR\ AM+FM\ SRS}}$ $[F(1,8) = 43.9, p < 0.0005]$, for $f_{cut\mathrm{AM}}$ = 50 Hz. For $f_{cut\mathrm{AM}}$ = 500 Hz, the same conclusion can be reported: the A-NMSE$_{\mathrm{AM\text{-}ONLY\ SRS}}$ is significantly greater than the A-NMSE$_{\mathrm{SV\ AM+FM\ SRS}}$ $[F(1,8) = 17.6, p < 0.005]$, and the A-NMSE$_{\mathrm{FR\ AM+FM\ SRS}}$ $[F(1,8) = 17.9, p < 0.005]$. However, no significant difference is revealed between the

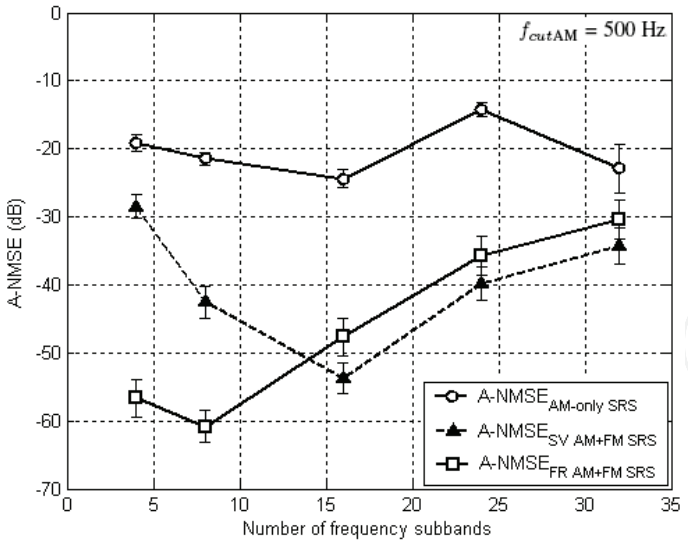


Fig. 6. The A-NMSE$_{\text{AM-ONLY SRS}}$ (solid-line + circle), A-NMSE$_{\text{SV AM+FM SRS}}$ (dashed-line + triangular), and A-NMSE$_{\text{FR AM+FM SRS}}$ (solid-line + square), calculated on 250 utterances (section III.A). The AM extraction lowpass filter cutoff frequency, $f_{cut\mathrm{AM}}$, equals 500 Hz. As in Fig. 5, the error bars indicate 95% of the confidence interval around each A-NMSE.

A-NMSE$_{\text{SV AM+FM SRS}}$ and the A-NMSE$_{\text{FR AM+FM SRS}}$, both with $f_{cut\mathrm{AM}}$ = 50 Hz $[F(1,8) = 1.8, p > 0.2]$, and $f_{cut\mathrm{AM}}$ = 500 Hz $[F(1,8) = 0.8, p > 0.35]$.

In addition, for every number of SRS frequency subbands, we can remark that A-NMSE$_{\text{AM-ONLY SRS}}$ is always the greatest amongst the three A-NMSEs. This gap reflects that the distortion of F0 information in the state-of-the-art cochlear implant acoustic simulation (AM-only SRS), is greater than the FAME-based SRS (SV AM+FM SRS and FR AM+FM SRS). This is a quantitative evidence which supports the report of (Zeng et al., 2005), about the low speaker recognition score (23%) of the cochlear implant users. This evidence supports also the fact that the normal hearing listeners, listening to the FAME-based SV AM+FM SRS, achieved significant better recognition scores, compared to when listening to the AM-only SRS, in the same speaker recognition task (Zeng et al., 2005). Further, we can remark that at low spectral resolution (4 and 8 subbands), the A-NMSE$_{\text{SV AM+FM SRS}}$ (dashed-line + triangular) is significantly greater than the A-NMSE$_{\text{FR AM+FM SRS}}$ (solid line + square), both for $f_{cut\mathrm{AM}}$ = 50 Hz $[F(1,2) = 15.2, p = 0.06]$ and for $f_{cut\mathrm{AM}}$ = 500 Hz $[F(1,2) = 10.01, p = 0.087]$. Even though the p-values in these two cases are slightly greater than 0.05, we can still state the latter conclusion since the error bars, which indicate 95% of the confidence interval around each A-NMSE, do not overlap (Cumming et al., 2007). This phenomenon suggests that the presence of rapidly varying FM components (above 400 Hz) in the FR AM+FM SRS, help in reducing the speech F0 distortion at low spectral resolution. However, at high SRS spectral resolution (16 subbands and above), the difference between the A-NMSE$_{\text{SV AM+FM SRS}}$ and the A-NMSE$_{\text{FR AM+FM SRS}}$ is not significant ($[F(1,4) = 0.08, p > 0.75]$, $f_{cut\mathrm{AM}}$ = 50 Hz, and $[F(1,4) = 0.37, p > 0.55]$, $f_{cut\mathrm{AM}}$ = 500 Hz). The presence of rapidly varying FM components in the SRS is therefore not significant to reduce the F0 distortion.

3.2.4 Examples

Fig. 7 and Fig. 8 show the spectrograms of a continuous speech utterance, original clean speech and the corresponding SRSs ($f_{cut\mathrm{AM}}$ = 50 Hz), spoken by a female speaker, selected from the set of 250 utterances as mentioned in section III.A. The blue curves in the spectrograms, estimated by using the Praat software (Boersma & Weenink, 2009), represent the speech F0 vectors, $\mathbf{f}_X, \mathbf{f}_Y$, extracted from the original clean speech and the synthesized SRS utterances. The range of the F0 is labeled on the right-hand side vertical axis of each spectrogram. Using the extracted F0 curve of the original clean speech as the reference, we can qualitatively remark that the F0 values are not correctly estimated from the AM-only SRS, whether we use 4 subbands [NMSE = -17.4 dB] or 16 subbands [NMSE = -13.1 dB]. Another concern is related to the extracted F0 values in the AM+FM SRS. The F0 information is less well estimated in the 4-subband SV AM+FM SRS (5(c)) [NMSE = -20 dB] than in the 4-subband FR AM+FM SRS (5(d)) [NMSE = -35.3 dB]. This remark is consistent with the fact that the rapidly varying FM components help in reducing the F0 distortion at low spectral resolution, as mentioned previously. This phenomenon does not happen with the 16-subband SV AM+FM SRS (6(c)) [NMSE = -72.6 dB] and the 16-subband FR AM+FM SRS (6(d)) [NMSE = -73.2 dB] where the F0 estimation is sufficiently good. Again, the fact that the rapidly varying FM components are not significant for reducing the F0 distortion at high SRS spectral resolution, compared to the slowly varying FM components, is typically verified in this example.

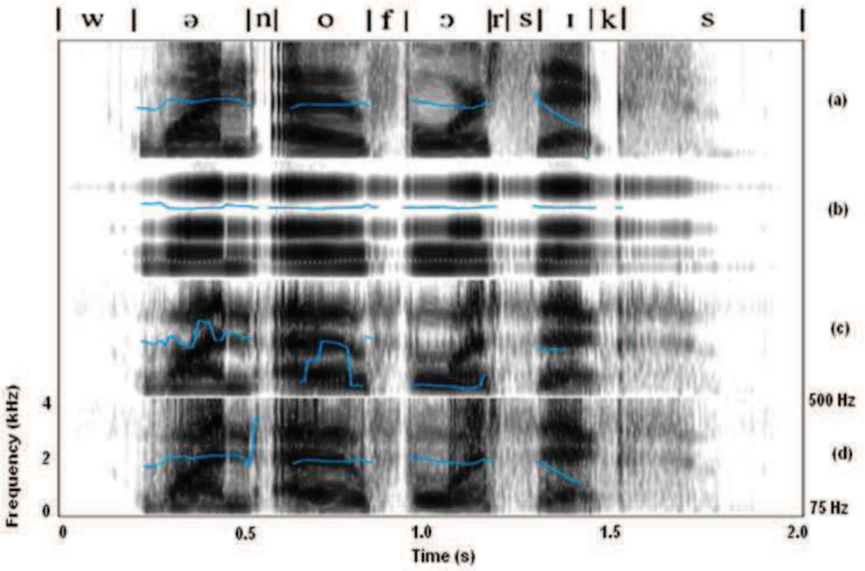


Fig. 7. Spectrograms of the original clean speech and the SRSs of the speech utterance "one oh four six", selected from the set of 250 utterances mentioned in section III.A, spoken by a female speaker; (a) original clean speech, (b) 4-subband AM-only SRS, (c) 4-subband SV AM+FM SRS, (d) 4-subband FR AM+FM SRS. The blue curves, estimated by Praat (Boersma & Weenink, 2009), represent the speech F0 vectors. The $f_{cut\mathrm{AM}}$ = 50 Hz and the F0 frequency range is [75 Hz - 500 Hz] (see the right-hand side vertical axis).

3.2.5 Concluding remarks

We have quantitatively analyzed the speech fundamental frequency, F0, in the cochlear implant-like spectrally reduced speech. The NMSE is calculated between the F0 values extracted from the original clean speech and those of the SRSs using algorithm 1, proposed in section III.B. NMSE analysis showed that amongst the three types of SRS studied in this chapter, the AM-only SRS is the SRS in which the speech F0 distortion is the greatest. The great distortion of F0 information in the state-of-the-art cochlear implant-like SRS, supports

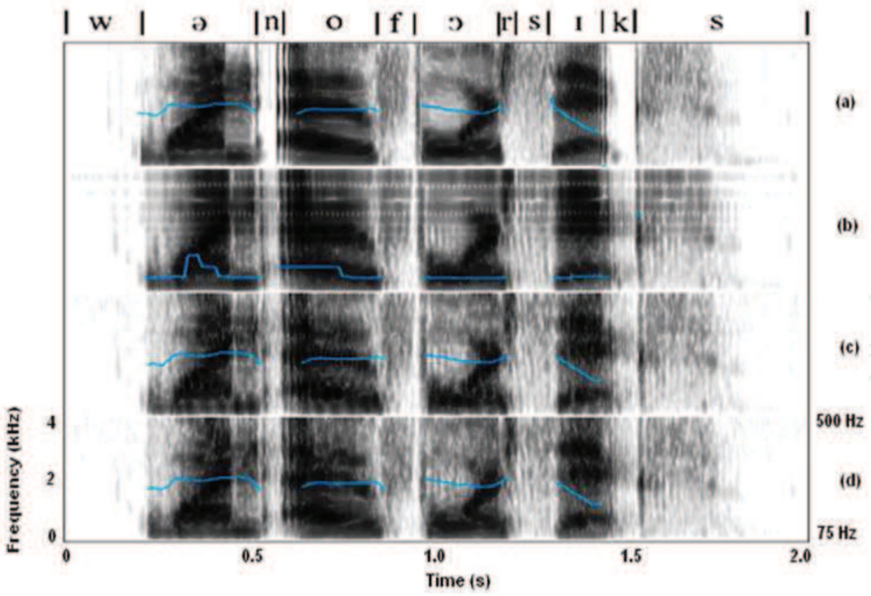


Fig. 8. Spectrograms of the original clean speech and the SRSs of the same utterance as in Fig. 8; (a) original clean speech, (b) 16-subband AM-only SRS, (c) 16-subband SV AM+FM SRS, (d) 16-subband FR AM+FM SRS. The $f_{cut\mathrm{AM}}$ = 50 Hz and the F0 frequency range is [75 Hz - 500 Hz].

the report of (Zeng et al., 2005): " [...] *current cochlear implant users can largely recognize what is said, but they cannot identify who say it.*" The FAME strategy (Nie et al., 2005), which proposes to extract the slowly varying FM components and integrates them in the cochlear implant, help in reducing the F0 distortion in the SV AM+FM SRS, compared to the AM-only SRS. However, at low spectral resolution (4 and 8 subbands), the rapidly varying FM components are beneficial to reduce the F0 distortion in the FR AM+FM SRS. At high SRS spectral resolution (16 subbands and above), there is no significant difference between the SV AM+FM SRS and the FR AM+FM SRS in terms of F0 distortion. The results obtained in this chapter might help improve the FAME strategy (Nie et al., 2005) for better performance in a speaker recognition task, by keeping the rapidly varying FM components when the SRS spectral resolution is low. Even though rapidly varying FM components cannot be perceived by cochlear implant listeners, their presence could improve speaker recognition performance of cochlear implant listeners. Obviously, further subjective studies are needed to confirm this suggestion. A similar remark can be found in a study of Chang, Bai and Zeng in which the authors, by a subjective study, have shown that the low-frequency sound component below 300 Hz, *"although unintelligible when presented alone, could improve the functional signal-to-noise ratio by 10-15 dB for speech recognition in noise when presented in combination with a cochlear implant simulation"* (Chang et al., 2006).

The speech F0 NMSE analysis performed on the cochlear implant-like SRS is a quantitative evidence supporting the speaker recognition subjective tests, performed in (Zeng et al., 2005). On the other hand, quantitative studies on other speech acoustic features could also be performed, in advance, on the acoustic simulation of cochlear implant (SRS) to orient subjective tests. The results of such quantitative analysis could be exploited to conduct subjective studies in cochlear implant research.

3.3 Automatic recognition of acoustic simulations of cochlear implants

3.3.1 Relevant speech spectral information for ASR

It is important to reduce speech signal variability, due to speech production (accent and dialect, speaking style, etc.) or environment (additive noise, microphone frequency response, etc.), in order to guarantee stable ASR performance. Therefore, in an ASR system, the speech analysis module is aimed at reducing the speech signal variability and extracting the ASR relevant spectral information into speech acoustic features. However, despite the speech variability reduction achieved by such standard speech signal analyses, ASR performance is still adversely affected by noise and other sources of acoustic variability (Raj & Stern, 2005). Since most standard speech processing analyses for ASR are performed in the spectral domain, it is natural to seek the relevant spectral information that is sufficient for ASR, in the speech signal.

In ASR based on Hidden Markov Models (HMMs) (Rabiner, 1989), one way to estimate the ASR relevant speech spectral information is to evaluate the ASR performance on spectrally reduced speech (SRS) signals when the acoustic models (the HMMs) are trained on a clean speech (full spectrum) database. As usual, the tested signals must not belong to the training database. Such an approach was first investigated by Barker and Cooke in (Barker & Cooke, 1997) where the authors *"consider how acoustic models trained on original clean speech can be adapted to cope with a particular form of spectral distortion, namely reduction of clean speech to sin-wave replicas"*. The ASR results were, however, not satisfactory in train-test unmatched conditions. The cepstral coding techniques are, following the authors, *"inappropriate for dealing with drastic alterations to the shape of the spectral profile caused by spectral reduction"* (Barker & Cooke, 1997).

The acoustic simulation of cochlear implant is a spectrally reduced transform of original speech (Shannon et al., 1995). This type of SRS should be appropriate to evaluate the relevant spectral information needed by an HMM-based ASR whose acoustic models are trained on a given clean speech database and which uses the Mel frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980) or the perceptual linear prediction (PLP) coefficients (Hermansky, 1990) as acoustic features. The rationale is twofold. On the one hand, cochlear implant-like SRS can be recognized by normal hearing listeners. The recognition scores then depend on the spectral resolution (or the number of frequency subbands) of the SRS (Shannon et al., 1995). Furthermore, human cochlear implant listeners relying on primarily temporal cues can achieve a high level of speech recognition in quiet environment (Zeng et al., 2005). The foregoing facts suggest that the cochlear implant-like SRS could contain sufficient information for human speech recognition, even though such an SRS is synthesized from the speech temporal envelopes only. On the other hand, in ASR, certain speech analyses, such as the Bark or Mel-scale warping of the frequency axis or the spectral amplitude compression, performed on the conventional speech acoustic features (MFCCs or PLP coefficients), derive from the model of the human auditory system. Such auditory-like analyses, which mimic the speech processing performed by the human auditory system, are basically aimed at reducing speech signal variability and emphasizing the most relevant spectral information for ASR (Morgan et al., 2004). As a result, the cochlear implant-like SRS should contain sufficient spectral information for ASR based on conventional acoustic features, such as the MFCCs or the PLP coefficients.

3.3.2 Experimental data

In this section, the SRSs are synthesized from 250 original utterances (see 3.2.2) by using the AM-only SRS synthesis algorithm described in section 3.1. With 5 values (4, 8, 16, 24 and 32) of the number of frequency subbands and 4 values (16, 50, 160 and 500 Hz) for the bandwidth of the subband temporal envelope (AM), we have thus 20 sets of synthesized AM-only SRS signals, each set contains 250 utterances. These sets will be used for the spectral distortion analyses and the ASR tests in the next sections.

3.3.3 Spectral distortion analysis

Given an original clean speech utterance x_i and the corresponding SRS utterance $\widehat{x}_i$ which is synthesized from x_i, assume that f_j and $\widehat{f}_j$ are the spectra of two speech frames of x_i and $\widehat{x}_i$, respectively. The spectral distortion between these two speech frames can be measured on their spectra $\mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j)$. A good spectral distortion measure should have the following properties (Nocerino et al., 1985):

1. $\mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j) \geq 0$, with equality when $f_j = \widehat{f}_j$;
2. $\mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j)$ should have a reasonable perceptual interpretation;
3. $\mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j)$ should be numerically tractable and easy to compute.

Amongst the available spectral distortion measures (Nocerino et al., 1985), the weighted slope metric (WSM) distortion measure (Klatt, 1982), which is a perceptually based distortion measure, satisfies well these three properties. This spectral distortion measure reflects the spectral slope difference near spectral peaks in a critical-band spectral representation (Nocerino et al., 1985). It is also shown that the WSM distortion measure correlates well with the perceptual data (Klatt, 1982). Further, the WSM distortion measure is one of the spectral distortion measures that gave the highest recognition score with a standard dynamic time warping (DTW) based, isolated word, speech recognizer (Nocerino et al., 1985). In this respect, we use the WSM distortion measure to assess the spectral distortion between two speech frames extracted from an original clean speech utterance and a synthesized SRS utterance, respectively. The mathematical formula of the WSM distortion measure has the form (Klatt, 1982; Nocerino et al., 1985)

$$\mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j) = k_E \left| E_{f_j} - E_{\widehat{f}_j} \right| + \sum_{q=1}^{Q} k_s(q) \left(s_{f_j}(q) - s_{\widehat{f}_j}(q) \right)^2 \tag{3}$$

where Q is the number of frequency subbands, k_E is a weighting coefficient on the absolute energy difference, $\left| E_{f_j} - E_{\widehat{f}_j} \right|$, between f_j and $\widehat{f}_j$, $k_s(q)$ is a weighting coefficient for the difference, $s_{f_j}(q) - s_{\widehat{f}_j}(q)$, between the two critical band spectral slopes of f_j and $\widehat{f}_j$ (Nocerino et al., 1985).

Henceforth, $\mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j)$ designates the WSM distortion measure. The length of the speech frames for the WSM distortion measure is 10 ms and the number of frequency subband $Q = 24$. The Matlab programs for calculating the WSM spectral distortion can be found at (Loizou, 1999b). The spectral distortion $\eta_{x_i,\widehat{x}_i}$ between two speech utterances x_i and $\widehat{x}_i$ can be defined as the mean of the WSM spectral distortion measures between the speech frames $\mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j)$:

$$\eta_{x_i,\widehat{x}_i} = \frac{1}{M}\sum_{j=1}^{M} \mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j) \tag{4}$$

where M is the number of speech frames in the speech utterances x_i and $\widehat{x}_i$. We define the overall spectral distortion $\bar{\eta}$ as the average of the $\eta_{x_i,\widehat{x}_i}$ calculated for all the $N = 250$ utterances in each set of SRS signals:

$$\bar{\eta} = \frac{1}{N}\sum_{i=1}^{N} \eta_{x_i,\widehat{x}_i} = \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M} \mathrm{d}_{x_i,\widehat{x}_i}(f_j,\widehat{f}_j) \tag{5}$$

We then calculate the overall spectral distortion $\bar{\eta}$ for all 20 sets of SRS utterances (see section 3.3.2). The values of $\bar{\eta}$ are illustrated in Fig. 9. The error bars in Fig. 9 represent the standard deviations of $\eta_{x_i,\widehat{x}_i}$. An ANOVA revealed no significant difference between the overall spectral distortion $\bar{\eta}$ when the bandwidth of the subband temporal envelopes are changed from 16 Hz to 500 Hz [F(3,16) = 0.64, p $>$ 0.5]. Generally, the overall spectral distortion $\bar{\eta}$ decreases when the number of frequency subbands of the SRS increases. Since the WSM distortion measure is perceptually based and is correlated with the perceptual data (Klatt, 1982), we can deduce that the value of $\bar{\eta}$ reflects more or less the level of perceptual distortion that the listeners have to deal with when listening to the SRS signals. In addition, we can remark that $\eta_{x_i,\widehat{x}_i}$ varies intensively when the number of frequency subbands of the SRS is small (4- and 8-subband SRS), since the standard deviations of $\eta_{x_i,\widehat{x}_i}$ are large at low spectral resolutions of the SRS (4 and 8 subbands).

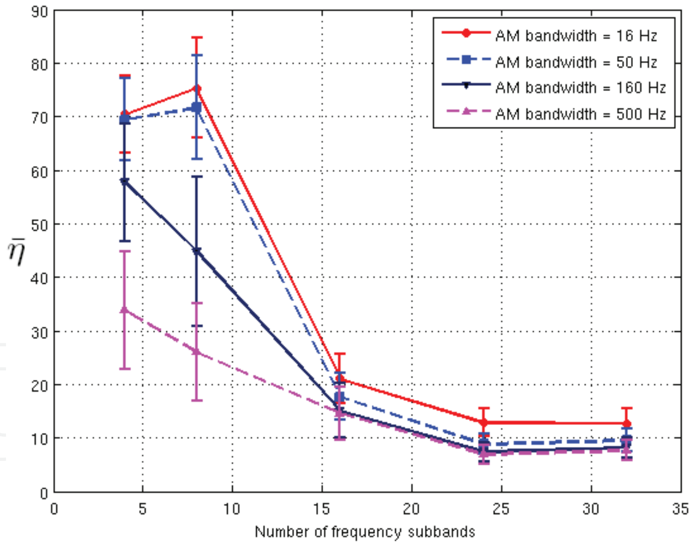


Fig. 9. Overall spectral distortion $\bar{\eta}$, which is defined as the average of the spectral distortion $\eta_{x_i,\widehat{x}_i}$, are calculated on all the 250 utterances of each set of SRS signals. There would be no overall spectral distortion if $\bar{\eta} = 0$. Error bars indicate standard deviations.

3.3.4 Automatic speech recognition results

We used the HTK speech recognition toolkit (Young et al., 2006) to train a speaker-independent HMM-based ASR system on the TI-digits speech database (Leonard, 1984). TI-digits is a large speech database of more than 25 thousand connected digits sequences spoken by over 300 men, women, and children. This speech database is widely used in the literature to assess ASR algorithms on small-vocabulary tasks (Barker et al., 2005; Cooke et al., 2001; Smit & Barnard, 2009). The data were collected in a quiet environment and digitized at 20 kHz. In this study, the data were downsampled to 8 kHz. The ASR system used a bigram language model and the acoustic models were the context-dependent three-state left-to-right triphone HMMs. These models were trained by using both MFCCs and PLP coefficients. The output observation distributions were modelled by Gaussian mixture models (GMMs) (Rabiner, 1989). In each state, the number of mixture components was 16. The feature vectors consist of 13 MFCCs or 13 PLP coefficients. For the MFCCs and the PLP coefficients calculation, the standard filterbank consisting of 26 filters was used (Young et al., 2006). The MFCCs and the PLPs coefficients were calculated from every Hamming windowed speech frame of 25 ms length and with an overlap of 15 ms between two adjacent frames. The first (delta) and second (acceleration) difference coefficients were appended to the static MFCCs and PLP coefficients to provide 39-dimensional feature vectors. This configuration for the feature vectors was used in both training and testing conditions. Next, the ASR tests were taken on the 20 sets of synthesized SRS signals. The recognition results with the MFCCs and PLP coefficients are shown in Table 1 and Table 2, respectively.

AM bandwidth	Number of frequency subbands				
	4	8	16	24	32
16 Hz	35.44	68.33	96.22	96.22	96.29
50 Hz	35.57	84.23	99.82	99.70	99.57
160 Hz	37.76	95.86	99.63	99.70	99.63
500 Hz	39.34	97.56	99.57	99.70	99.63

Table 1. ASR word accuracies (in %) computed on the 250 SRSs synthesized from 5 values for the number of frequency subbands and 4 values for the bandwidth of the subband temporal envelopes (AM). The ASR system was trained on the TI-digits clean speech training database. The speech feature vectors were MFCC-based. The ASR word accuracy computed on the 250 original clean speech utterances is 99.76%.

AM bandwidth	Number of frequency subbands				
	4	8	16	24	32
16 Hz	35.32	87.27	97.99	98.66	98.48
50 Hz	35.57	98.36	99.57	99.57	99.57
160 Hz	35.69	98.96	99.57	99.63	99.70
500 Hz	41.17	99.15	99.57	99.63	99.57

Table 2. ASR word accuracies (in %) computed on the 250 SRSs synthesized from 5 values for the number of frequency subbands and 4 values for the bandwidth of the subband temporal envelopes (AM). The ASR system was trained on the TI-digits clean speech training database. The speech feature vectors were PLP-based. The ASR word accuracy computed on the 250 original clean speech utterances is 99.70%.

For the ASR results with MFCC-based speech feature vectors, an ANOVA performed on the lines of Table I revealed that changing the bandwidth of the subband temporal envelopes had no significant effect in terms of ASR word accuracy [F(3,16) = 0.11, $p > 0.95$]. However, another ANOVA performed on the columns of Table I indicated that changing the number of frequency subbands had a significant effect across all tests [F(4,15) = 73.9, p < 0.001]. Protected *t*-tests (or Least Significant Difference (LSD) tests) (Keren & Lewis, 1993) were thus performed and showed that the 8-subband SRS yielded significant better ASR word accuracies compared to the 4-subband SRS [$t_{obs(4,8)} = 11.23 > t_{crit(15)} = 4.07, \alpha = 0.001$], where $t_{obs(4,8)}$ is the protected *t*-test value calculated between the 4-subband SRS and the 8-subband SRS word accuracies, $t_{crit(15)}$ is the critical value at the desired α level for 15 degrees of freedom. In contrast, no significant difference was revealed amongst the 16, 24, and 32-subband SRSs in terms of ASR word accuracy [$t_{obs(16,24)} \approx t_{obs(24,32)} \approx t_{obs(16,32)} \approx 0 < t_{crit(15)} = 2.13, \alpha = 0.05$]. The ASR word accuracies of each SRS in this group are significant better than those of the 8-subband SRS [$t_{obs(8,16)} = 2.79, t_{obs(8,24)} = 2.80, t_{obs(8,32)} = 2.78, t_{crit(15)} = 2.13, \alpha = 0.05$].

For the ASR results with PLP-based speech feature vectors, the same statistical analyses had been performed on the ASR word accuracies in Table 2 and the same conclusions would be revealed. In summary, increasing the number of the SRS frequency subbands from 4 to 16 made significant improvement in terms of ASR word accuracy. Interestingly, the 16, 24, and 32-subband SRSs could achieve an ASR word accuracy comparable to that attained with the original clean speech signals (the ASR word accuracies computed on the 250 original clean speech utterances with MFCC-based and PLP-based speech feature vectors were 99.76% and 99.70%, respectively).

3.3.5 Conclusions and perspectives

We have investigated the automatic recognition of cochlear implant-like SRS, which is synthesized from subband temporal envelopes of the original clean speech. The MFCCs and the PLP coefficients were used as speech features and the ASR system, which is speaker-independent and HMM-based, was trained on the original clean speech training database of TI-digits. It was shown that changing the bandwidth of the subband temporal envelopes had no significant effect on the ASR system word accuracy. In addition, increasing the number of frequency subbands of the SRS from 4 to 16 improved significantly the performance of the ASR system. However, there was no significant difference amongst the 16, 24, and 32-subband SRS in terms of ASR word accuracy. It was possible to achieve an ASR word accuracy as good as that attained with the original clean speech by using SRS with 16, 24, or 32 frequency subbands and by using both MFCC-based and PLP-based speech features.

The MFCCs or the PLP coefficients, along with the delta and acceleration coefficients, were concatenated together in the acoustic feature vector and were used in both the training and the testing conditions. The results presented in this chapter suggest that SRS with 16, 24, and 32 subbands contain sufficient spectral information for speech recognition with HMM-based ASR system using the MFCCs or the PLP coefficients along with the delta and acceleration coefficients. The SRS and original clean speech are quite different in the signal domain since the SRS is synthesized from original clean speech temporal envelopes, only. The spectral distortion analysis, based on the WSM distortion measure, showed that there are significant spectral distortions in the synthesized SRS compared to the original clean speech signal. Despite of these spectral distortions which are induced by the SRS synthesis, and although the ASR acoustic models are trained on the clean speech training database (TI-digits), the ASR word accuracy, computed on original clean speech signals of a testing set of TI-digits, is still maintained with the synthesized SRS of 16, 24, and 32 subbands.

Therefore, the 16-, 24-, and 32-subband SRS models might lead to the design of new acoustic features and suggest new speech models, for ASR, that could be robust to noise and other sources of acoustic variability. In this respect, performing ASR with HMMs trained on SRS could assess the relevance of SRS as a model for speech recognition. In addition, the fact that cochlear implant-like SRS is synthesized from speech temporal envelopes, only, might help in reducing ASR irrelevant spectral information due to environment. In (Do et al., 2012), we perform the ASR experiments with the noise contaminated speech signal. The SRS are used in both training and testing conditions. That is, the ASR system is trained on the SRS which was synthesized from the training original clean speech. This system is be used to recognize the SRS which is synthesized from noisy contaminated speech signals. Experimental results show that the use of SRS helps in improving significantly the ASR performance in the recognition of noisy speech (Do et al., 2012).

4. General conclusions and discussions

In this chapter, we have introduced the applications of acoustic simulations of cochlear implants in human and machine hearing research. We have reviewed the conventional applications of the acoustic simulations of cochlear implants in (1) determining the number of independent channels needed to achieve high levels of speech understanding, and in (2) simulating the effect of electrodes insertion depth on speech identification accuracy. In addition, we have introduced novel applications of acoustic simulations of cochlear implants, abbreviated as SRS, both in cochlear implant research and machine hearing research (or automatic speech recognition). Conventional as well as recently proposed synthesis algorithms have been used to synthesize the SRS signals (Nie et al., 2005; Shannon et al., 1995). These novel applications show that the acoustic simulation of cochlear implant is an important tool, not only for cochlear implant research but also for other hearing-related applications, namely noise robust ASR (Do et al., 2012).

However, for any studies performed on acoustic simulations of cochlear implants, the results should be interpreted with cautions when applied on human subjects who are cochlear implant users. In fact, the results could be slightly different due to the changes that occur in their auditory system following the severe and profound deafness. In this respect, results of the studies performed with acoustic simulations should be used as general indicators to conduct useful subjective studies. Authentic conclusions could be afterward interpreted from the results of studies performed on human subjects. On the other hand, in section 3.3, it was shown that changing the bandwidth of the subband temporal envelopes has no significant effect on the ASR word accuracies. In addition, changing the number of subband from 8 to 16 increases significantly the ASR word accuracies. Physically, these facts suggest that ASR word accuracies are more affected by the change of spectral information distributing on the whole spectral band rather than the change of local spectral information in each subband.

5. Acknowledgements

Some parts of the work in this chapter have been performed when the author was with Télécom Bretagne, Brest, France, in collaboration with Dominique Pastor and André Goalic.

6. References

Barker, J. & Cooke, M. (1997). Modelling the recognition of spectrally reduced speech, *Proc. ISCA Eurospeech 1997, September 22 - 25, Rhodes, Greece*, pp. 2127–2130.

Barker, J., Cooke, M. & Ellis, D. (2005). Decoding speech in the presence of other sources, *Speech Communication* 45(1): 5–25.

Baskent, D. & Shannon, R. V. (2003). Speech recognition under conditions of frequency-place compression and expansion, *J. Acoust. Soc. Am.* Vol. 113(4): 2064–2076.

Boersma, P. (2004). Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound, *Proc. of Institut of Phonetic Sciences, University of Amsterdam*, Vol. 17, pp. 97–110.

Boersma, P. & Weenink, D. (2009). Praat, Doing Phonetic by Computer (Version 5.1.12)[computer program], *http://www.praat.org* .

Chang, J. E., Bai, J. Y. & Zeng, F.-G. (2006). Unintelligible Low-frequency Sound Enhances Simulated Cochlear-implant Speech Recognition in Noise, *IEEE Trans. Biomed. Eng.* 53(12): 2598–2601.

Chatterjee, M. & Peng, S.-C. (2008). Processing F0 with Cochlear Implant: Modulation Frequency Discrimination and Speech Intonation Recognition, *Hearing Research* 235(1-2): 143–156.

Cooke, M., Green, P., Josifovski, L. & Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Communication* 34(3): 267–285.

Cumming, G., Fidler, F. & Vaux, D. L. (2007). Error Bars in Experimental Biology, *The Journal of Cell Biology* 177(1): 7–11.

Davis, S. B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences, *IEEE Trans. Acoustics, Speech, Signal Processing* 28(4): 357–366.

Do, C.-T., Pastor, D. & Goalic, A. (2010a). On normalized MSE analysis of speech fundamental frequency in the cochlear implant-like spectrally reduced speech, *IEEE Trans. on Biomedical Engineering* Vol. 57(No. 3): 572–577.

Do, C.-T., Pastor, D. & Goalic, A. (2010b). On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR, *IEEE Trans. on Audio, Speech and Language Processing* Vol. 18(No. 5): 2993–2996.

Do, C.-T., Pastor, D. & Goalic, A. (2012). A novel framework for noise robust ASR using cochlear implant-like spectrally reduced speech, *Speech Communication* 54(1): 119–133.

Do, C.-T., Pastor, D., Le Lan, G. & Goalic, A. (2010). Recognizing cochlear implant-like spectrally reduced speech with HMM-based ASR: experiments with MFCCs and PLP coefficients, *Proc. Interspeech 2010, September 26 - 30, Makuhari, Japan*.

Dorman, M. F., Loizou, P. C. & Rainey, D. (1997a). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs, *J. Acoust. Soc. Am.* 102(4): 2403–2411.

Dorman, M., Loizou, P. C. & Rainey, D. (1997b). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding, *J. Acoust. Soc. Am.* Vol. 102(No. 5): 2993–2996.

Dudley, H. (1939). Remarking speech, *J. Acoust. Soc. Am.* (11): 169–177.

Faulkner, A., Rosen, S. & Wilkinson, L. (1997). Effects of the number of channels and speech-to-noise ratio on rate of connected discourse tracking through a simulated cochlear implant speech processor, *Ear & Hearing* pp. 431–438.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species - 29 years later, *J. Acoust. Soc. Am.* Vol. 87(No. 6): 2592–2605.

Gunawan, T. S. & Ambikairajah, E. (2004). Speech Enhancement using Temporal Masking and Fractional Bark Gammatone Filters, *Proc. 10th Australian International Conference on Speech Science & Technology, December 8 - 10, Sydney, Australia*, pp. 420–425.

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech, *J. Acoust. Soc. Am.* 87(4): 1738–1752.

Keren, G. & Lewis, C. (1993). *A handbook for data analysis in the behavioral sciences: statistical issues*, Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey Hove & London.

Klatt, D. H. (1982). Prediction of perceived phonetic distance from critical band spectra: A fisrt step, *Proc. IEEE ICASSP 1982, May 03 - 05, Paris, France*, Vol. 2, pp. 1278–1281.

Kubin, G. & Kleijn, W. B. (1999). On Speech Coding in a Perceptual Domain, *Proc. IEEE ICASSP, March 15 - 19, Phoenix, AZ, USA*, Vol. 1, pp. 205–208.

Leonard, R. (1984). A Database for Speaker-independent Digit Recognition, *Proc. IEEE ICASSP, March 19 - 21, San Diego, USA*, Vol. 9, pp. 328–331.

Loizou, P. C. (1999a). Introduction to cochlear implants, *IEEE Engineering in Medicine and Biology Magazine* 18(1): 32–42.

Loizou, P. C. (1999b). COLEA: A Matlab software tool for speech analysis [Computer program] [Online]. Available: http://www.utdallas.edu/ loizou/speech/colea.htm.

Loizou, P. C., Dorman, M. & Tu, Z. (1999). On the number of channels needed to understand speech, *J. Acoust. Soc. Am.* Vol. 106(No. 4): 2097–2103.

Morgan, N., Bourlard, H. & Hermansky, H. (2004). Automatic speech recognition: an auditory perspective, *Speech processing in the auditory system (S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay, Eds.)* SPRINGER pp. 309–338.

Nie, K., Stickney, G. & Zeng, F.-G. (2005). Encoding frequency modulation to improve cochlear implant performance in noise, *IEEE Trans. on Biomedical Engineering* Vol. 52(No. 1): 64–73.

Nocerino, N., Soong, F. K., Rabiner, L. R. & Klatt, D. H. (1985). Comparative study of several distortion measures for speech recognition, *Speech Communication* 4(4): 317–331.

Patterson, R. D., Robinson, K., Holdsworth, J., Zhang, C. & Allerhand, M. H. (1992). Complex sounds and auditory images, *Auditory physiology and perception (Y. Cazals, L. Demany, and K. Horner, eds.)* OXFORD, PERGAMON pp. 429–446.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE* 77(2): 257–286.

Raj, B. & Stern, R. M. (2005). Missing-feature approaches in speech recognition, *IEEE Signal Processing Magazine* 22(5): 101–116.

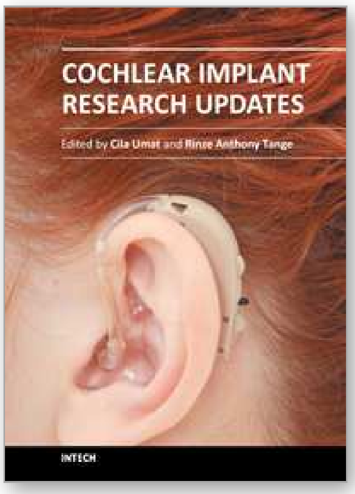
Shannon, R. V., Fu, Q.-J. & Ganvil, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation, *Acta Otolaryngol. Suppl.* (552): 50–54.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues, *Science* Vol. 270(No. 5234): 303–304.

Smit, W. & Barnard, E. (2009). Continuous speech recognition with sparse coding, *Computer Speech and Language* 23(2): 200–219.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2006). *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, Cambridge, UK.

Zeng, F.-G., Nie, K., Stickney, G., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C. & Cao, K. (2005). Speech recognition with amplitude and frequency modulations, *PNAS* Vol. 102(No. 7): 2293–2298.


COCHLEAR IMPLANT
RESEARCH UPDATES
Edited by Cila Umat and Rinze Anthony Tange
INTECH