Speech Communication 91 (2017) 17–27

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Speech intelligibility improvement in car noise environment by voice transformation

Karan Nathwani[a,*], Gaël Richard[a], Bertrand David[a], Pierre Prablanc[b], Vincent Roussarie[b]

[a] LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
[b] PSA Peugeot Citroën, Chemin de Gisy, 78943 Vélizy-Villacoublay

ARTICLE INFO

Article history:
Received 27 July 2016
Revised 3 February 2017
Accepted 25 April 2017
Available online 5 May 2017

Keywords:
Speech intelligibility
Car noise environment
Hearing in noise test
Voice transformation
Lombard speech

ABSTRACT

The typical application targeted by this work is the intelligibility improvement of speech messages when rendered in car noise environment (radio, message alerts,...). The main idea of this work is to transform the original speech to "Lombard" speech or more precisely to simulate some of the strategies followed by humans to render their speech clearer when they are surrounded by noise. Three main effects are considered in this work, namely non uniform-time scale modification, formant shifting and a combination of these modifications along with energy redistribution between speech regions. All effects are studied with specific transformations for voiced and unvoiced segments. The proposed modifications are then evaluated by means of subjective and objective tests. The results of these tests conducted with normal hearing and impaired listeners demonstrate the potential of the selected transformations for voice intelligibility improvement.

1. Introduction

Speech Intelligibility usually refers to a measure of the effectiveness of understanding speech. It is used to evaluate telecommunications systems performances, to characterize some acoustical properties of conference rooms or to evaluate the level of understanding of patients for medical purposes.

It is well known that intelligibility is affected by the presence of background noise which can mask crucial portions of the speech content. It is also widely acknowledged that humans succeed to enhance the audibility of their voice by means of a number of non-linear effects which are often gathered under the term of Lombard effect (Junqua, 1993; Lombard, 1911; Van Summers et al., 1988; Lu and Cooke, 2008; Cooke et al., 2013). In fact, humans do modify their speech in the presence of noise in such a way to enhance the acoustic contrasts between their speech and the background noise. It is for example shown in Garnier and Henrich (2014) that besides a straightforward strategy of speaking louder, the speakers did alter other vocal characteristics such as for example the center frequency of the first formant or the modulation of the fundamental frequency on voiced segments.

For many applications including speech recognition or telephony in adverse conditions it is necessary to reduce the impact of the surrounding noise and to mitigate channel effects. This explains the vast literature on speech enhancement and dereverberation Hodoshima et al. (2002); Arai et al. (2002). However, improving speech quality does not necessarily improve speech intelligibility (Kim and Loizou, 2010; Hu and Loizou, 2007) and factors explaining why speech quality is not directly linked to speech intelligibility are studied but not yet well understood (Loizou and Kim, 2011). This may be explained by the fact that speech enhancement algorithms maximize a cost function which may not be well correlated with speech intelligibility.

In this work, we focus on a different application, namely the improvement of the intelligibility of a spoken message (possibly originally uttered in a quiet environment) when rendered in a noisy environment. A typical application concerns the intelligibility improvement of speech messages in the car environment (radio, alert messages, telephony,...). The main idea of this work is to transform the original speech to Lombard speech or more precisely to simulate some of the strategies followed by humans to render their speech more audible when they are surrounded by noise. In such in-car applications, it is always possible to increase the volume of the speech sources, due to the limitation of the audio system dynamics and moreover for hearing deficient persons it is not necessarily suitable. This justifies to seek appropriate voice transformations without significantly changing the sound loudness.

Various acoustic features of Lombard speech could be investigated such as the increase of the fundamental frequencies on specific phonemes, formant center frequencies shifting, sound

* Corresponding author.
E-mail address: karan046@gmail.com (K. Nathwani).

http://dx.doi.org/10.1016/j.specom.2017.04.007

energy redistribution from voiced to unvoiced sounds or vice-versa, speech rate modification or spectral tilting (Lombard, 1911; Van Summers et al., 1988; Lu and Cooke, 2008; Cooke et al., 2013). However, to keep the perceptual study tractable, we have only considered three of these effects in this work, namely formant shifting, time-scale modification and energy redistribution (ER) with specific transformations for voiced and unvoiced segments. The proposed modifications have been evaluated by means of subjective and objective tests. The results of these tests conducted with normal hearing and impaired listeners demonstrate the potential of the selected transformations for voice intelligibility improvement.

The remainder of the paper is organized as follows. A brief review of the Lombard effect and voice transformation techniques is given in Section 2. We then describe in Section 3 the proposed phonetically-dependant time scale modification. We explain in Section 4 the process for achieving high quality formant shifting for voiced segments and the integration of the different transformations into a single system (duration scaling, formant shifting and energy redistribution) is described in Section 5. Section 6 is dedicated to performance evaluation. A brief conclusion and some perspectives are then suggested in Section 7.

2. Lombard speech and voice transformation

There exists a rich literature on finding acoustic-phonetic correlates of intelligibility. For example, it was found that female voices have a higher intelligibility (Hazan and Markham, 2004; Barker and Cooke, 2007), that intelligibility is increased when the formant vocalic triangle is enlarged Bond and Moore (1994); Ferguson and Kewley-Port (2002) or when, in presence of noise, the energy ratio between consonant and vowels is stronger Hazan and Simpson (1998). Other studies aimed at understanding the non-linear effects of Lombard speech and what are the differences with normal speech. These studies have shown that speech uttered in the presence of background noise compared to normal speech may exhibit a lower speech rate, an increase of its fundamental frequency, a spectral flattening (or increase of the higher frequency or some frequency regions such as around 3kHz), (Lu and Cooke, 2008; Steeneken and Hansen, 1999; Garnier and Henrich, 2014). Nevertheless, some results remain controversial and for example the role of the increase of fundamental frequency in Lombard speech is not consensual (Lu and Cooke, 2009).

The emergence of efficient voice transformation techniques would permit to transform natural speech into Lombard speech if all factors were well understood. Indeed, voice transformation is an active field of research which has witnessed a large number of novel approaches especially in the last decade (Stylianou, 2009; Machado and Queiroz, 2010). The methods may follow a speech production model and directly alter the model parameters (vocal tract and glottal source parameters). At the opposite, they may rely on a signal model as in Harmonic plus Noise models or parametric spectral estimation (Vincent et al., 2010). There also exist a number of hybrid methods exploiting some aspects of speech production and signal modelling including linear predictive models (Rao and Yegnanarayana, 2006), cepstral analysis or the STRAIGHT model (Kawahara, 1997).

For example, efficient algorithms exist for speech rate or pitch contour modification (Moulines and Laroche, 1995; Laures and Bunton, 2003). Similarly, a number of voice transformation approaches have been introduced to change the spectral content or spectral envelope. Such approaches are typically based on a conventionally learning phase for finding the statistical differences between the parameters of two voices using Hidden Markov Models, Gaussian Mixture Model or neural networks (Nurminen et al., 2006; Desai et al., 2009) or for learning a transformation function (weighted linear interpolations and bilinear models (Zhang et al., 2008)).

However, less work is done in the context of improving speech intelligibility in noise environment by synthesizing some aspects of the Lombard effect. For example, in Skowronski and Harris (2006), energy redistribution between voiced and unvoiced segments is implemented by moving signal energy to targeted regions of relatively high information content which are important for intelligibility. The boosted regions are originally of low energy and therefore redistributing the energies to such regions will increase the intelligibility. In another work, speech energy is redistributed over time and frequency according to a perceptual distortion measure (which is based on a spectro-temporal auditory model) under noise environment (Taal et al., 2014). In Taal and Jensen (2013), a linear time-invariant filter is designed in order to improve speech understanding in noise by maximizing the speech intelligibility index (SII) under the constraint that the speech energy is held constant. However, the noise used in the aforementioned methods was either white noise or speech shaped noise.

InNathwani et al. (2016), we aimed at improving speech intelligibility in car environment by applying formant modifications. It was, in particular, shown that shifting the central frequency of the lower formants away from the region of noise resulted in higher intelligibility despite the audible degradation of the speech quality. One of the outcome of this preliminary study was that the tested voice transformation was promising but limited. It was also found that the artifacts introduced by the voice transformation were particularly detrimental for intelligibility.

In this paper, we propose an extension of this preliminary work by incorporating three main effects such as non uniform-time scale modification, smoothed shifting of formants for voiced segments, and a combination of these modifications along with energy redistribution. As highlighted in the literature, we have designed different modifications for voiced and unvoiced segments. Time scaling modification is applied with a different scaling factor for voiced and unvoiced segments. Formant shifting is only applied on voiced segments and is dynamically smoothed to limit the pitfalls of the previous approach. Finally, the third modification includes the combination of non uniform time scaling, smoothed shifting of formants for voiced parts and energy redistribution between voiced and unvoiced segments to synthesize the Lombard speech. These different modifications are further described in the following sections.

3. Speech rate modification

Slowing down the recorded speech helps transcribing recorded notes and practicing spoken languages, and blind people with developed hearing sense may prefer accelerated speech when listening to recorded audio books to save time. Speech rate is obviously an important parameter in speech intelligibility. In this work we explore the potential impact of lowering the speech rate with different scaling factor for voiced and unvoiced segments Kupryjanow and Czyzewski (2012), Erogul and Karagoz (1998), Yang et al. (2008).

3.1. Algorithm for non uniform-Time scaling modification (NU-TSM)

The motivation for using different scaling factor comes from the fact that lengthening of the vowels, consonants and phone-transitions have different impact on speech intelligibility (Kupryjanow and Czyzewski, 2012; Erogul and Karagoz, 1998) compared to the uniform time scaling. In Kupryjanow and Czyzewski (2012), it is in particular shown that time scaling the vowel by a higher factor than for the consonants, while keeping the phone-transitions unchanged improves intelligibility.

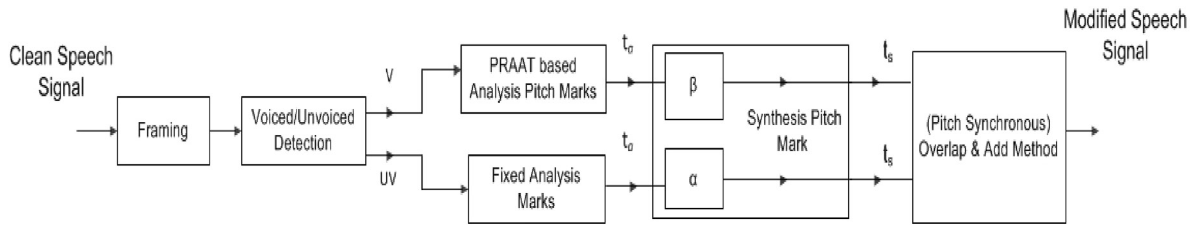

Fig. 1. Flow diagram for the proposed non uniform-time scale modification.

Fig. 1 shows the flow diagram for the non uniform - time scale modification based on different scaling factor for voiced and unvoiced segments. The clean speech signal is first segmented into successive overlapping frames. Thereafter, the voiced and unvoiced (V/UV) decision is made framewise. The V/UV knowledge is then used to obtain the analysis pitch marks (t_a) from the clean speech signal. The synthesis pitch marks (t_s) are then obtained from the analysis pitch marks by using different scaling factor for V/UV segments. Finally short time analysis signals are then combined based on streams of synthesis pitch marks using the pitch synchronous overlap and add (PSOLA) method to obtain time scaled signal.

The algorithmic steps involved in time scale modification with non uniform scaling factor for voiced and unvoiced segments are as follows.

1. **Input**: Clean Speech Signal, α (unvoiced scaling factor) and β (voiced scaling factor).
2. **Framing**: The clean speech signal is segmented into frames by using 25 ms Hanning window.
3. **Voiced and Unvoiced (V/UV) detection**: The segmented speech frame is then passed to V/UV segment detector. In this work, the yet another algorithm for pitch tracking (YAAPT) algorithm (Zahorian and Hu, 2008) which uses the combination of spectral and temporal information to track the best candidate for fundamental frequencies, is used for V/UV decision. However, the normalized low frequency energy ratio (NLFER) is used as initial V/UV detector for pitch tracking in YAAPT as detailed in Zahorian and Hu (2008).
4. **Analysis Pitch Mark (APM) selection**: The analysis pitch marks for voiced segments are then estimated using the pitch values obtained from the previous step. The procedure to obtain pitch marks is based on the PRAAT pitch marking algorithm (Kotnik et al., 2006; Hagmüller and Kubin, 2006). PRAAT algorithm is based on the auto-correlation method to perform acoustic periodicity detection. The implementation for computing the analysis pitch marks using PRAAT is as follows. The first pitch mark is obtained as the maximum in the interval $t_0 - \frac{T}{2}$ to $t_0 + \frac{T}{2}$. Here, t_0 corresponds to the middle of the speech excerpt and T is the pitch period at t_0. The previous (resp. next) pitch mark of t_0 is searched in the interval $t_0 - 1.2T$ to $t_0 - 0.8T$ (resp. $t_0 + 0.8T$ to $t_0 + 1.2T$). The exact locations are given by the maximum in this interval. This process is repeated until all pitch marks are found. For unvoiced segments, the analysis marks are equally spaced according to the analysis window length.
5. **Synthesis Pitch Mark selection**: Next step is to obtain synthesis pitch marks with different scaling factors for voiced and unvoiced regions. The first synthesis pitch mark is taken equal to the first analysis pitch mark. If the next frame is unvoiced, t_s for the current frame ($(m)^{th}$) is obtained as

$$t_s(m) = t_s(m-1) + \alpha(t_a(m) - t_a(m-1)) \tag{1}$$

On the other hand, if the frame is voiced, t_s for the current frame ($(m)^{th}$) is obtained as

$$t_s(m) = t_s(m-1) + \beta(t_a(m) - t_a(m-1)) \tag{2}$$

6. **Synthesis**: Finally, the non uniform-time scaled signal is obtained by combining the analysis waveforms synchronized on the stream of synthesis pitch marks using the PSOLA method.
7. **Normalization**: The variance of the synthesized signal is finally normalized to the original signal.
8. **Output**: Non uniform time scaled signal.

4. Formant shifting

The formant shifting procedure described in Nathwani et al. (2016) improved the intelligibility for a majority of listeners in high speed car noise but with however a high standard deviation in the statistical performance. This variability in the results can be explained by artifacts introduced by the voice transformation algorithm. If at very low SNRs, these artifacts are inaudible (e.g. they are masked by the background noise), they become annoying and detrimental to intelligibility at high SNRs. However, it may be noted that the formant shift method (Nathwani et al., 2016) remain rather small limiting the possibilities of vowel changes. Furthermore, it was noticed that artifacts due to sound transformation was the most detrimental factor for intelligibility decrease. Additionally, It should also be noted that the HINT protocol (Vaillancourt et al., 2005) (used in this work for subjective evaluation) using whole sentences is probably less sensitive to localised (potential) vowel changes than other types of intelligibility tests based on short meaningless phonetic sequences.

Hence, we propose in this work two extensions to our preliminary study. First, we here apply formant shifting to voiced segments only. Indeed, the concept of formant shifting does not seem to be effective for unvoiced sounds and leaving the unvoiced component unaltered helps to better preserve naturalness of the transformed speech (Amano-Kusumoto and Hosom, 2011; Moon and Lindblom, 1994). Second, we have observed that artifacts are mostly due to sudden changes in formant trajectories. We then introduce a smoothing step which softens the altered formant trajectories and therefore limits the pitfalls of the previous approach.

4.1. Algorithm for smoothed shifting of formants for voiced segments (SSFV)

The formant shifting procedure described in Nathwani et al. (2016) is optimized for a typical car noise and is not adaptive to variable noise characteristics. However, the formant shifting procedure could lead to audible artifacts when applied too aggressively. The smoothing technique has been added to avoid abrupt jumps in formant shifts from a frame to another and thus to lessen those artifacts. Fig. 2 shows the block diagram for smoothed shifting of formants for voiced segments based voice transformation technique. The clean speech signal is first segmented into successive frames. Each speech frame is then identified as either voiced or unvoiced segment. The auto regressive (AR) modelling is then applied on each frame to get the linear prediction (LP) coefficients. From LP coefficients, the poles and formant location are computed. The formants of the voiced segments are then finally shifted away

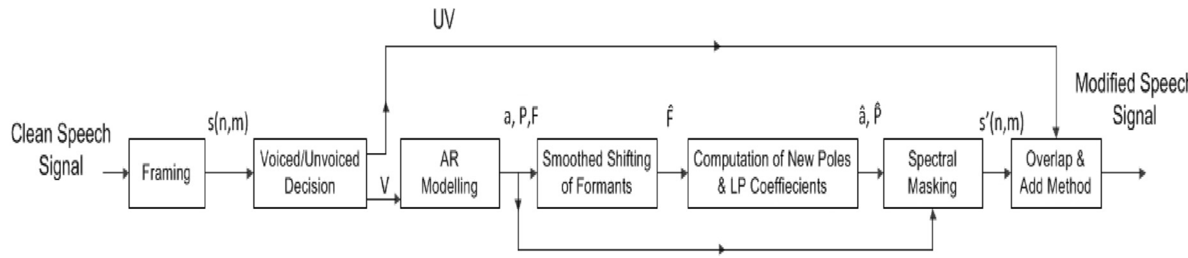

Fig. 2. Block diagram illustrating the proposed smoothed shifting of formant for voiced segments.

from the region of noise. The amount of formant shift in Hertz for each speech frame is characterized by a delta function $\delta(F)$ (see below). The smoothing step consists of low-pass filtering the successive frequency shift values using a simple exponential smoothing model (Brown, 1959). Finally, an updated set of LP coefficients are obtained from the shifted poles and served to synthesis step (e.g. Spectral Masking and Overlap and Add) to obtain the modified speech signal.

The detailed algorithmic steps of the proposed smoothed shifting of formants for voiced segments are given below.

1. **Input**: Clean speech signal, smoothing factor (ζ).
2. **Framing**: The clean speech signal is segmented into shorter frames using a Hanning window of 25 ms duration.
3. **Voiced and Unvoiced (V/UV) decision**: The voiced and unvoiced decision is then made on each frame using the YAAPT algorithm (Zahorian and Hu, 2008). The unvoiced segments are unaltered during the process. However, the LPC coefficients are computed from the voiced segments.
4. **AR modeling**: An AR model is a powerful front end tool to process speech signal. In AR model, a speech frame signal $s(n, m)$ can be expressed in terms of a p^{th} order linear predictor (Rabiner and Schafer, 1978; Nathwani et al., 2016). Here, n and m correspond to speech sample and short time frame indices respectively. The order of linear predictor (p) is equal to 12.
5. **Poles and formants computation**: The LP filter $A(f, m)$ is then computed from LP coefficients ($a_k(m)$ of the m^{th} frame) as

$$A(f,m) = 1 + \sum_{k=1}^{p} a_k(m) e^{-j2\pi fk} \tag{3}$$

The poles $P(k, m)$ and formant frequencies $F(k, m)$ are then estimated as the roots of the LP filter $A(f, m)$ and the angle of estimated poles respectively. Here k and f correspond to the formant frequency index and STFT frequency bin index respectively.
6. **Smoothed shifting of formants**: The formants obtained from voiced segments in previous steps are then shifted upwards by an amount specified by a delta function $\delta(F)$. The formants from unvoiced segments are not shifted during the process. The delta function $\delta(F)$ used in this case is shown in Fig. 3. It may be noted that the delta function shape should depend on the noise statistics. In this work, some instances of typical car noises were used to design a simple piecewise linear shape for the delta function so that the formants are shifted away from the noise region. As described in Nathwani et al. (2016) we have chosen the different shapes of the delta function based on the best PESQ and SII scores for a given car noise. This is obviously suboptimal in real conditions with noises of variable spectral characteristics. However, we rather aim at demonstrating the potential of the proposed approach and the design of an optimal noise adaptive delta function is beyond the scope of this paper and left for future research. Once the value of the shift $\rho(k, m)$ is obtained for the k-th formant frequency by applying the delta function such that $\rho(k,m) = \delta(F(k,m))$, the new

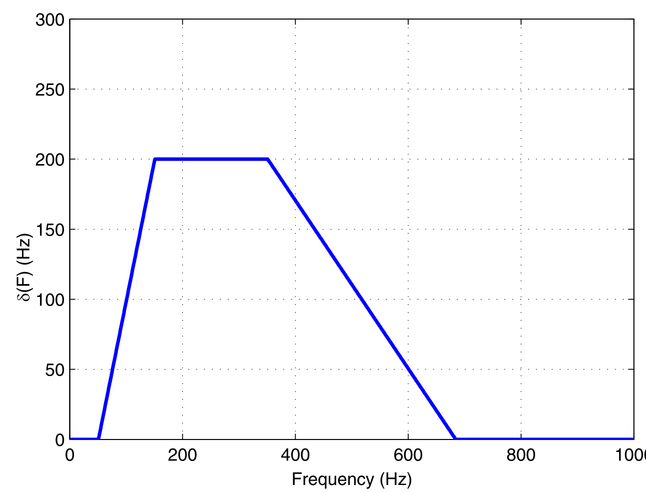

Fig. 3. Delta function used for the proposed smoothed shifting of formants for voiced segments.

formant frequency $\hat{F}(k,m)$ for the m^{th} frame is obtained in the following manner:

$$\hat{F}(k,m) = F(k,m) + \Delta(k,m) \tag{4}$$

where the smoothed shifting value ($\Delta(k, m)$) is obtained by smoothing $\rho(k, m)$ across the time frames using a simple exponential model (Brown, 1959). Hence

$$\Delta(k,m) = \zeta\rho(k,m) + (1-\zeta)\rho(k,m-1) \tag{5}$$

where ζ denotes a positive factor such as $0 < \zeta < 1$.
7. **Computation of new poles and LP coefficients**: The new poles $\hat{P}(k,m)$ are then computed from the estimated new formant frequency $\hat{F}(k,m)$ as

$$\hat{P}(k,m) = Be^{j2\pi\hat{F}(k,m)} \tag{6}$$

Here, B is the amplitude of the original poles. The modified LP coefficients are then easily obtained from the new filter built for the new poles $\hat{P}(k,m)$.
8. **Spectral masking**: A direct synthesis using the modified LP coefficients results in good speech quality on average but with some rare but annoying localised artifacts. To reduce these artifacts which are mostly due to phase incoherence, the magnitude of the original speech spectrum is modified in order to match the modified LP spectrum while the original phase spectrum is kept.
Given $S(f,m) = |S(f,m)|e^{j\phi(f,m)}$, the short time Fourier transform of an original speech frame $s(n, m)$ and $|A(f, m)|$ is the module of the LP spectrum of $s(n, m)$, the modified speech frame spectrum $S'(f, m)$ is then obtained as

$$S'(f,m) = |S(f,m)|.|A'(f,m)|/|A(f,m)|e^{j\phi(f,m)} \tag{7}$$

Here, $A'(f, m)$ is the STFT of the modified LP coefficients and $\phi(f, m)$ is the short time phase spectrum. The modified speech

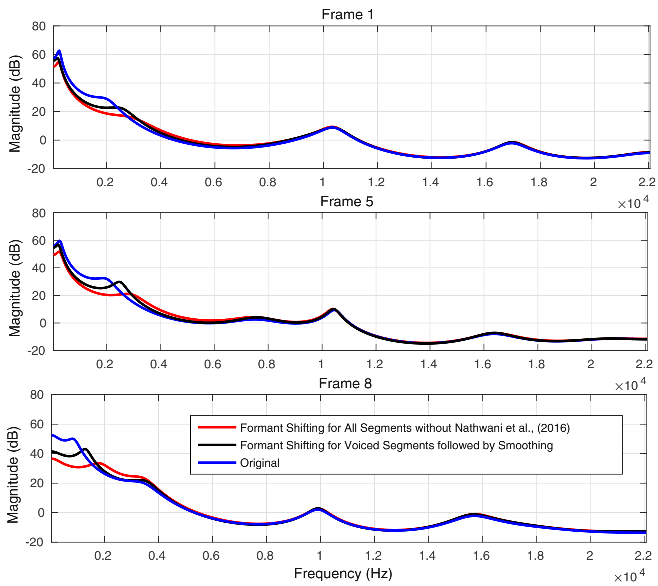

Fig. 4. Filter spectrum of original signal, formant shifted signal (Nathwani et al., 2016) and signal obtained by smoothed shifting of formants for voiced segments, in particular frames.

frame is then obtained by inverse Fourier transform and its overall energy is equalized to the energy of the unprocessed frame.

9. **Overlap and add (OLA) method**: Finally, the modified signal is obtained using classic Overlap and Add synthesis.
10. **Output**: Modified speech signal.

4.2. Significance of smoothing in artifacts reduction

In Nathwani et al. (2016), it was particularly shown that shifting the central frequency of the lower formants away from the noise region resulted in higher intelligibility despite the audible degradation of the speech quality. These degradations are possibly due to the artifacts generated during the voice transformation. One of the main reasons of such artifacts was due to the sudden changes in formant trajectories across the frames.

In Fig. 4, an example is shown for the artifacts introduced by the voice transformation algorithm proposed in Nathwani et al. (2016). In order to illustrate the effect of smoothing the altered formant trajectories, the filter spectrum for the original signal, formant shifted signal for voiced segments followed by smoothing and formant shifted signal (FS) without smoothing (Nathwani et al., 2016), are presented for particular voiced speech frames. It can be seen from Fig. 4 that the lower formants of FS method is shifted too aggressively compared to the original signal causing disturbance in the naturalness of the signal and degradation of the audio quality. However in the SSFV filter spectrum, formant shifting is only applied on the voiced segments leaving the unvoiced segments unaltered to preserve the naturalness of the signal. This is followed by dynamically smoothing to soften the altered formants trajectories and therefore limit the pitfalls of the previous approach (Nathwani et al., 2016). Thus, the SSFV filter spectrum has been able to shift the spectrum away from the region of noise without causing significant artifacts in the spectrum.

In an another attempt to justify the significance of the smoothing in artifacts reduction, the delta function values obtained before and after smoothing for first and second formant frequencies are shown in Fig. 5 for a particular sentence.

It can be seen from Fig. 5 that $\Delta(F)$ and $\delta(F)$ is zero for unvoiced segments as there is no shifting performed for such segments as explained in Section 4. On the other hand, the delta function values obtained before ($\delta(F)$) and after smoothing ($\Delta(F)$) should have some non zero coefficients for voiced segments, depending on which formant frequencies lie inside the delta function shape. However, during transition from voiced to unvoiced segments, the first unvoiced segment is likely affected due to smoothing used in $\Delta(F)$ compared to $\delta(F)$. Thus, Fig. 5 indicates that the smoothing step softens the altered formant trajectories when $\Delta(F)$ is used in formant shifting instead of $\delta(F)$.

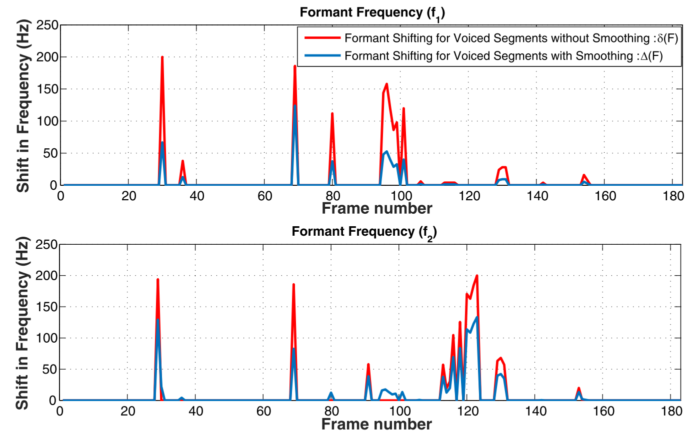

Fig. 5. Delta function values used in proposed formant shifting for voiced segments with and without smoothing are illustrated for first and second formant frequencies across the time frames.

Indeed, in Fig. 4 of the manuscript, we have displayed the modifications by showing the magnitude spectra of some voiced speech frames for the original signal, the formant shifted signal without smoothing Nathwani et al. (2016) and formant shifted signal with smoothing. For most unvoiced segments, it is not desirable to perform formant shifts. As a result, in the SSFV approach we only transformed voiced frames and use a smoothing step to guaranty smooth transitions in time. To illustrate this, we have shown the successive speech spectra of two segments containing a transition from voiced (resp. unvoiced) to unvoiced (resp. voiced) frames in Fig. 6.

It has been observed that in general, during transition from unvoiced to voiced segment and vice-versa, the filter spectrum of SSFV at end points of unvoiced segments are most likely to be different from the original filter spectrum. This could be due to smoothing effect which results in very small non zero values for $\Delta(F)$ for transition unvoiced segments. The similar observation can be seen for SSFV filter spectrum from Fig. 6.

5. Combined speech modifications

In this section, we propose to combine several modifications namely, the non-uniform time scaling, smoothed shifting of formants for voiced segments and energy redistribution.

5.1. Energy redistribution

Energy redistribution as introduced in Skowronski and Harris (2006) is a rather simple modification which automatically increases the intelligibility of speech in noisy environment while preserving the overall signal power and naturalness of the original speech. More precisely, the rationale behind the energy redistribution modification is to boost unvoiced segments and to reduce voiced segments with the constraint to limit sound harshness. This in turn mimics the Lombard effect by accentuating phonetic contrast. Originally, the energy of the signal is "moved" to targeted regions of relatively high information content which are important

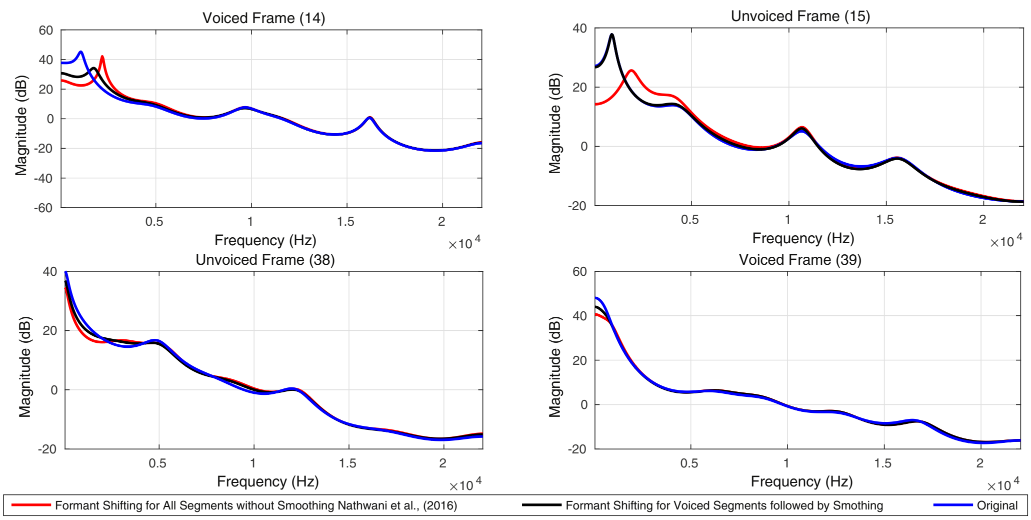

Fig. 6. Filter spectrum for consecutive voiced and unvoiced frames.

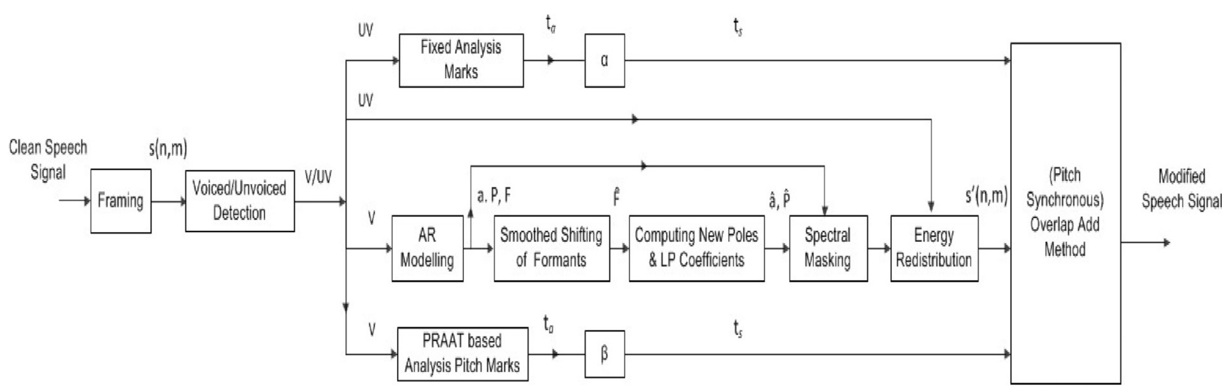

Fig. 7. Flow diagram of the proposed combined duration scaling, smoothed shifting of formants for voiced segments and energy redistribution modification.

for intelligibility. The boosted regions are originally of low energy and therefore redistributing the energy to such regions increases intelligibility while preserving the naturalness of the signal.

Here, energy redistribution is done between voiced and unvoiced segments only. Unvoiced regions typically have less power than voiced regions and are more easily obscured by noise in the listener's environment. By boosting unvoiced regions, the energy of the unvoiced speech is raised above the noise resulting in intelligibility increment. The transition between voiced and unvoiced gain factors is finally smoothed by a 20 ms linear interpolation. The word utterance is then scaled by a normalizing gain factor such that the modified word energy is the same as the original word energy. It may also be noted that the scaling factor used to boost the unvoiced segments is selected in such a way that naturalness of the signals is preserved.

5.2. Algorithm for combined speech modifications

Fig. 7 illustrates the block diagram of the proposed combined speech modification obtained by Fusion of non uniform-Time scaling, Smoothed shifting of Formants for voiced segments and Energy redistribution (FTSFE) between voiced and unvoiced segments.

The algorithmic steps for achieving the combined speech modifications are as follows.

1. **Input**: Clean speech signal, λ (unvoiced gain factor), μ (voiced gain factor), smoothing factor (ζ), α (unvoiced scaling factor) and β (voiced scaling factor).
2. The clean speech signal is first segmented into successive frames by using a Hanning window of 25 ms. Thereafter, the voiced and unvoiced decision is made framewise using the YAAPT algorithm.
3. The analysis pitch marks and fixed analysis marks (t_a) are then estimated from the voiced and unvoiced segments respectively. However, it may also be noted that PRAAT pitch marking algorithm (Kotnik et al., 2006; Hagmüller and Kubin, 2006) is used to compute pitch marks for voiced segments. On the other hand, analysis marks for unvoiced segments are obtained at fixed analysis window. This is followed by synthesis pitch marks (t_s) computation from t_a. The detailed procedure of analysis and synthesis pitch marks selection can be seen from Section 3.1.
4. The smoothed shifting of formant algorithm described in Section 4 (AR modelling, smoothed shifting of formants for voiced segments, new poles computing and spectral masking) is then applied on voiced segments to obtain the shifted formant speech frame.
5. The smooth shifted formant speech frame is then multiplied by either a voiced (μ) or unvoiced (λ) gain factor depending on the voiced or unvoiced segment respectively. This step smoothly redistributes the energy between voiced and unvoiced segments.
6. The modified speech signal is then obtained by combining the modified (smoothed formant shifted) speech frames synchronized on the streams of synthesis pitch marks t_s. Thus, the pitch synchronous overlap-add method is used for this purpose.

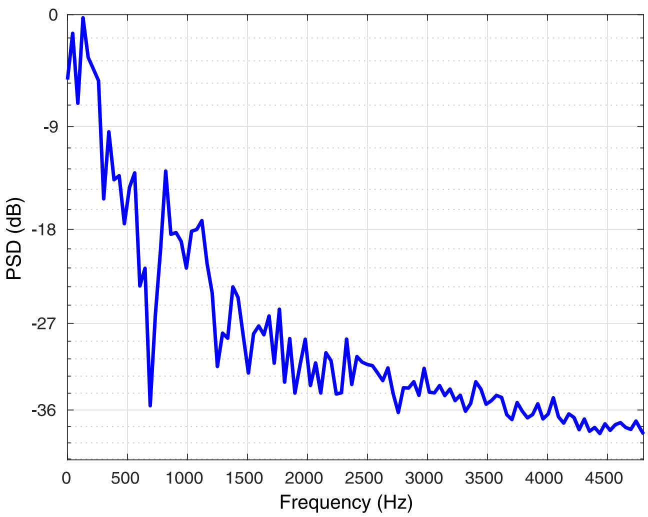

Fig. 8. Power Spectrum Density (PSD) of the car noise recorded at 130 km/h.

7. The overall energy of the modified signal is then normalized to the original signal.
8. **Output**: Modified speech signal.

6. Performance evaluation

The impact of the proposed modifications on intelligibility is evaluated using subjective and objective evaluation at different SNRs. The subjective test is performed based on the hearing in noise test (HINT) protocol (Nilsson et al., 1994). The objective measures used for evaluating intelligibility are speech intelligibility index (SII) Taal et al. (2010), perceptual evaluation of speech quality (PESQ) Rix et al. (2001), log likelihood ratio (LLR), weighted spectral slope (WSS) Loizou (2013) and mutual information (MI) Taghia and Martin (2014). Spectrographic analysis is also performed to test its coherence with other evaluations.

6.1. Development of the database and material

The French lists for HINT were adapted from the English version in Vaillancourt et al. (2005). Hence, 5 lists of 20 sentences used for the test were taken from an audiometry CD recording (CND, 2015). The 4 modifications (ER, NU-TSM, SSFV, FTSFE) along with the reference clean signal (NM) were applied to those lists. The car noise was recorded using a Head Acoustics dummy head in a mid-size car at 130 km/h steady speed. The power density spectrum of the noise is illustrated in Fig. 8. It can be seen that it contains few energies/frequencies above 1000 Hz with a constant decrease of energies with frequency. Thus, the positive shift of the delta function in formant shifting should increase the SNR of the shifted formants.

Finally the sentences obtained from 4 modifications along with the clean reference signal are mixed with the car noise recording at various SNRs (which are selected empirically). The mix was presented under Sennheiser HD650 headphones and played from a Head Acoustics Digital Equalizer (PEQ V). The level of the noise is set at 67 dB speech perception level (SPL) as it was the level in the car during the recording and only the level of the speech is varying across the experiments. Speech perception level seeks to understand how human listeners recognize speech sounds and use this information to understand spoken language.

6.2. Parameters selection for different modifications

In the duration scaling modification (NU-TSM), the voiced segments are scaled by $\beta = 1.4$ and unvoiced segments are scaled by $\alpha = 1.2$. In the context of SSFV modification, the smoothing factor ζ is equal to 0.66. It is observed that this smoothing factor value softens the altered formant trajectories, while preserving the naturalness of the signal. In fact, the selection of the smoothing factor is highly dependent on the altered formant trajectories which in turn is dependant on noise statistics. The unvoiced (λ) and voiced (μ) gain factor used for energy redistribution are 1.4 and 0.9 respectively. However, it may be noted that the same parameters are used in FTSFE modification as used in individual modifications.

The general strategy to select the aforementioned parameters are based on best PESQ and SII scores for different modifications on the HINT database. In this work, PESQ and SII scores are obtained between the synthesized signal (obtained from ER, NU-TSM, SSFV and FTSFE) and the synthesized signal added with noise at different SNRs. These PESQ and SII scores are also compared with the PESQ and SII scores of the "“no modification” case. We also make sure that modifications should be done in such a way that the naturalness of the signal should be preserved. This is ensured by not allowing the PESQ scores between the clean speech and synthesized speech to go less than 3.

6.3. Objective evaluation

6.3.1. Evaluation based on PESQ, LLR, WSS and SII measures

PESQ analyzes the speech signal sample-by-sample after the temporal alignment of corresponding excerpts of the synthesized signal (SS) and the synthesized signal added with noise (SSN). PESQ principally models mean opinion score (MOS) results that cover a scale from 1 (bad) to 5 (excellent). WSS is a distance measure which computes the weighted difference between the spectral slopes of SS and SSN in each frequency band. LLR is a LPC-based measure which finds the spectral envelope difference between the SS and SSN (Ma et al., 2009). SII model (Taal et al., 2010) basically calculates the average amount of speech information available to a listener. The value of the SII varies from 0 (completely unintelligible) to 1 (perfect intelligibility).

Table 1 shows the mean scores for all the objective measures at various SNRs for different modifications. These mean objective scores are computed on all the sentences of the database. In general, a method having higher SII, PESQ scores and lower LLR, WSS scores is supposed to lead to higher intelligibility (Loizou, 2013) for the modified speech compared to the original speech when played in noise. The objective scores for NU-TSM have not been reported herein since NU-TSM cannot be properly assessed by objective measures. This may be due to the needed alignment process (required in PESQ, SII etc.) which will likely compensate the time-scaling. Although it is not appropriate to evaluate the effect of time-scaling on intelligibility with these objective measures, it is possible to assess other approaches which combine several modifications. We have then included objective evaluations for FTSFE as it is the combination of different effects (SSFV, ER and NU-TSM).

It can be seen from Table 1 that higher SII and PESQ scores along with lower LLR and WSS scores are observed at various SNRs for the modification SSFV over NM. This indicates the significance of SSFV in intelligibility improvement under high speed car noise. In comparison to ER (Skowronski and Harris, 2006), SSFV has shown better objective scores at most of the SNR values. The third modification FTSFE has been an adequate compromise between the two modifications NU-TSM and SSFV by incorporating the properties of individual modifications. It can also be seen from the Table 1 that the FTSFE has best PESQ, WSS and LLR scores compared to other modifications. However, FTSFE has approximately similar SII scores to SSFV modification. This indicates that the effect of SSFV and ER modifications have compensated the time alignment effect in FTSFE objective scores. Additionally, we

Table 1
Mean objective scores for all 5 conditions using SII, PESQ, LLR and WSS measures at different SNRs. Here, **NM**: No Modification, **ER**: Energy Redistribution, **NU-TSM**: Non Uniform-Time Scale Modification, **SSFV**: smoothed shifting of formants for voiced segments and **FTSFE**: Fusion of Time Scale, Smoothed shifting of Formants for voiced segments and Energy Redistribution Modification.

Method	SNR=−26				SNR = −14				SNR = −8				SNR = 0				SNR = 10			
	SII	PESQ	LLR	WSS	SII	PESQ	LLR	WSS	SII	PESQ	LLR	WSS	SII	PESQ	LLR	WSS	SII	PESQ	LLR	WSS
NM	0.39	1.26	1.37	70.43	0.63	1.91	0.61	64.18	0.76	2.31	0.42	55.48	**0.89**	2.94	0.26	40.96	**0.98**	3.73	0.14	23.29
ER	0.37	1.41	0.96	71.94	0.61	2.04	0.58	64.47	0.75	2.45	0.40	55.14	**0.89**	3.06	0.23	40.05	0.97	3.80	**0.10**	21.69
SSFV	**0.42**	1.49	1.31	68.70	**0.64**	2.02	0.56	62.15	**0.77**	2.42	0.38	53.85	**0.89**	3.04	0.24	39.90	**0.98**	3.79	0.13	22.74
FTSFE	0.40	**1.57**	**0.92**	**67.76**	0.63	**2.19**	**0.53**	**60.27**	0.76	**2.59**	**0.37**	**51.60**	**0.89**	**3.22**	0.22	**37.79**	0.97	**3.90**	0.11	**21.13**
SSFV+ER	0.40	1.50	0.94	69.70	0.62	2.12	0.54	62.23	0.75	2.51	0.37	53.41	**0.89**	3.12	**0.21**	39.17	0.97	3.84	**0.10**	21.51

Table 2
Mean MI scores for all modifications at different SNRs.

Method	SNR = −26		SNR = −14		SNR = −8		SNR = 0		SNR = 10	
	MIP	MIR	MIP	MIR	MIP	MIR	MIP	MIR	MIP	MIR
NM	2.24	0.05	10.20	0.24	20.68	0.49	39.26	0.92	63.27	1.49
ER	2.46	0.06	10.87	0.26	21.31	0.51	39.51	0.94	63.30	1.51
NU-TSM	**3.74**	**0.09**	11.57	0.30	21.50	0.55	39.02	1.00	61.87	1.59
SSFV	2.59	0.06	12.23	0.27	22.51	0.53	40.50	0.95	**63.66**	1.50
FTSFE	3.00	0.08	**13.36**	**0.35**	**23.64**	**0.61**	**40.79**	**1.06**	63.34	**1.64**
SSFV+ER	2.76	0.06	12.45	0.29	22.71	0.54	40.36	0.96	63.41	1.51

have also obtained the scores for SSFV+ER to analyse its impact on intelligibility and compare with FTSFE. It is observed that the performance of SSFV+ER lies between FTSFE and SSFV at various SNRs. Hence, it can be concluded that the combined effect, whether it is FTSFE or SSFV+ER, has shown improvement in objective scores compared to corresponding individual modifications. However on increasing the SNR to further high values, it is observed that the difference in intelligibility between no modification and all modifications vanishes. This indicates that there is no need to modify the original signal at high SNRs.

6.3.2. Evaluation based on mutual information measure (Taghia and Martin, 2014)

Mutual information (MI) is an another measure used for evaluating the speech intelligibility improvement in noise. MI is a measure of dependence between the two random variables that account for higher order statistics, and hence to consider dependencies beyond the conventional second order statistics (Taghia and Martin, 2014; Taghia et al., 2012). To estimate the MI scores, input signals are first transformed into 15 subbands by using a 1/3 octave band filter bank. Thereafter, the mutual information between the amplitude envelopes of the reference and the test signals is estimated per subband to evaluate auditory perception (Taghia and Martin, 2014). Here, reference signal correspond to synthesized signal obtained from different modifications (including NM) and test signal is the corresponding synthesized signal added with noise.

Table 2 presents the normalized mean mutual information scores in percentage (MIP) and instrumental intelligibility scores before normalization (MIR) (Taghia and Martin, 2014; Taghia et al., 2012) computed on complete database. The proposed modification FTSFE illustrates higher MIP and MIR scores over other modifications at various SNRs indicating its significance for speech intelligibility tasks. In general, the higher the mutual information scores, the stronger the dependency between the reference and the test signal. This in turn leads to a higher intelligibility improvement for the corresponding modification. SSFV+ER modification shows the second best performance for mutual information scores after FTSFE. However, it is interesting to observe that MIR scores for NU-TSM are slightly better than SSFV at various SNRs. Additionally, as the SNR increases to very high values, the MIP and MIR scores are approximately similar for all modifications. Thus, the Tables 1 and 2 indicate that inclusion of several modification is beneficial for intelligibility improvement under car noise environment.

6.4. Subjective evaluation

6.4.1. HINT protocol and participants for subjective evaluation

A slightly modified Hearing In Noise Test (HINT) (Vaillancourt et al., 2005) is used for measuring the enhancement of intelligibility provided by the different speech treatments. A speech reception threshold (SRT) is obtained for each 5 conditions. A total of 31 native French speaking male and female subjects participated to the test. Out of 31 subjects, 19 subjects were screened out normal hearing (mean besides 20 dB HL over 0.5–6 kHz) and remaining 12 subjects were hearing impaired. Their ages ranged from 21 to 59 years with a mean age of 35.

The subjects are asked to listen to a sentence and to repeat aloud what they hear. The first sentence is presented at a level below the SRT, usually at -30 dB. Then the SNR is increased by 2 dB steps until it is repeated correctly. The subsequent sentences are presented once each (in order to avoid training effect) at a level depending on the correct repetition of the preceding sentence. If it is repeated correctly, presentation level is attenuated by 2 dB, otherwise it is increased by 2 dB. For each condition, a 20 sentences list is presented in a random order. There is one list per condition. The presentation order of the 5 conditions is balanced over the participants as well as the presentation of the lists. The experimenter compares the listener's response to a text version to determine whether it is correct or not. Small variations in the sentences are allowed as specified in (Nilsson et al., 1994). The level of the noise is fixed to 67 dB SPL as it was the level in the car during the recording, only the level of the speech is varying between successive sound examples Vaillancourt et al. (2005).

6.4.2. Experimental results

In this section, speech reception thresholds (SRT) for each condition/modification is compared to the reference no modification (NM) condition. SRTs have been averaged over the last fifteen SNR obtained with HINT. The relative threshold (RT) (in dB) is obtained by taking the difference between the SRT of NM and SRT of each modification. A positive relative threshold indicates an improvement in intelligibility for the particular modification.

Table 3 illustrates the number of participants showing positive and negative RT (who gets benefit or not) from each modification.

Table 3
Number of participants having positive and negative relative thresholds for different modifications.

Methods	Total sum of RT ($TSRT$)		Different combinations of RT from −3 to 3 dB						
	$TSRT \leq -1$	$TSRT \geq 1$	$RT \leq -3$	$-3 < RT \leq -2$	$-2 < RT \leq -1$	$-1 < RT < 1$	$1 \leq RT < 2$	$2 \leq RT < 3$	$RT \geq 3$
ER	5	13	1	2	2	13	5	6	2
NU-TSM	8	10	1	2	5	13	4	5	1
SSFV	6	15	**0**	2	4	10	**7**	2	**6**
FTSFE	**1**	**23**	**0**	**1**	**0**	7	**7**	**10**	**6**

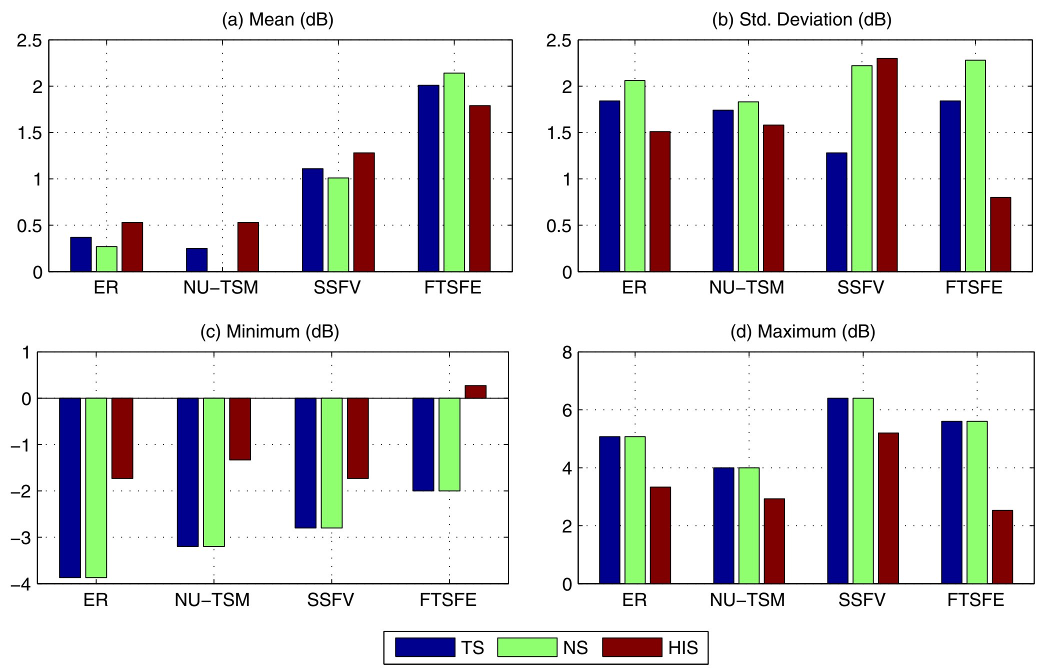

Fig. 9. Statistical analysis of the RT obtained from subjective evaluation.

We have considered that results with RT between −1 and 1 dB are not significant and hence are not further discussed below. In general, a method which gathers the highest and the lowest number of participants with positive and negative relative threshold respectively, is considered to be most useful for the intelligibility improvement task. The following conclusions which can be drawn from Table 3 are explained below.

- Out of total 31 subjects, the modifications ER, NU-TSM, SSFV and FTSFE have 13, 10, 15 and 23 subjects with RT significantly exceeding 1 dB respectively.
- Only one participant has $RT \leq -1$ dB for FTSFE modification compared to at least 5 participants for the other modifications.
- On the extreme cases of RT, the FTSFE and SSFV modifications have 0 and 6 participants each with $RT \leq -3$ and $RT \geq 3$ dB respectively, which is the best in comparison to other modifications.
- In addition, the modifications SSFV and FTSFE gather the greatest number of participants with $1 \leq RT < 2$ dB.

Hence in a nutshell, the proposed modification FTSFE has been most promising in the intelligibility improvement task in the high speed car noise by incorporating the properties of individual modifications. Additionally, the SSFV modification has shown second best performance after FTSFE.

Fig. 9 gives a statistical analysis of RT obtained from the subjective evaluation using total 31 subjects (TS), 19 normal subjects (NS) and 12 hearing impaired subjects (HIS). It can be seen from Fig. 9 that ER and NU-TSM do not provide significant mean improvement in intelligibility. On the contrary, SSFV and FTSFE show greater mean improvement in intelligibility for all subjects (including normal and impaired) with the highest maximum mean value and lowest minimum value is observed for SSFV and FTSFE respectively. Additionally, the standard deviation of the FTSFE modification indicates that it is possible to get a greater improvement (5.6 dB for maximum value) and a very low probability to have a possible degradation of the intelligibility (due to one subject).

When comparing the statistical analysis of HIS and NS populations, it is observed that the mean scores for ER, NU-TSM and SSFV are higher with lower standard deviation for HIS population in comparison to NS population. Although, the maximum mean values for all modifications are higher for NS compared to HIS population. On the other hand, the mean score for FTSFE modification under HIS population is lower than NS population. However, the least standard deviation and positive minimum mean value are observed for HIS in comparison to NS, when FTSFE modification is used.

In addition, we have run an ANOVA analysis along with a Duncan multiple range test which have highlighted that for impaired listeners the difference between SSFV and FTSF is not significative but that they both bring significant improvement compared to NM. For normal listeners, these statistic analyses have shown that the improvement brought by FTSFE is significant compared to all other approaches.

These findings indicates the significance of the proposed modifications in intelligibility improvement under high speed car noise for both normal and impaired listeners. Some of the French sound examples used in the experiments are accessible at http://perso.telecom-paristech.fr/~bedavid/karan/index.html. We have also presented some English sound examples on the webpage. However,

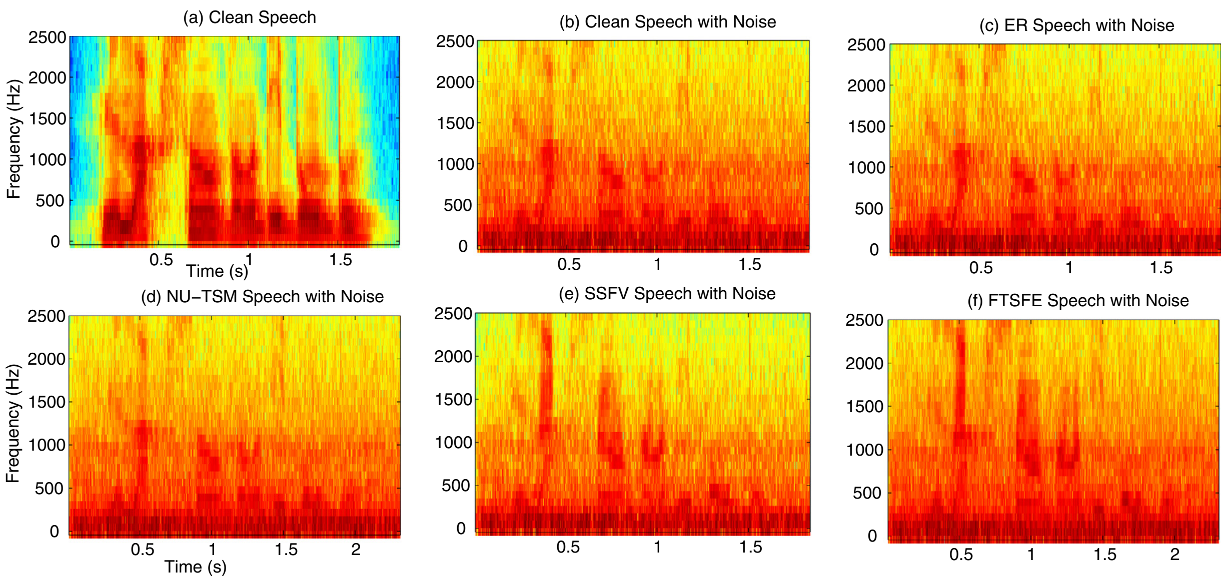

Fig. 10. Spectrograms of different modifications in the presence of noise at −8 dB SNR.

these English sound examples have not been used in the subjective and objective evaluations.

6.5. Spectrograhic analysis

This section deals with the spectrograhic analysis obtained from different modifications at SNR equal to -8 dB along with the clean speech spectrogram (Fig. 10 (a)). It may be noted that the speech signals are sampled at 44.1 KHz but their spectrograms are only displayed for frequencies up to 2.5 KHz since most speech modifications are below 1 KHz. It can be clearly seen from Fig. 10 (b) that the noise has masked low frequency spectrum of the clean speech completely, resulting in a significant loss of formants visibility which are crucial for intelligibility. The ER spectrogram in Fig. 10 (c) has strengthened the unvoiced segments which are obscured due to the addition of noise. This results in slight improvement of speech formants visibility in noise compared to NM. Similarly, the improvement in formants visibility is observed from NU-TSM spectrogram (Fig. 10 (d)) due to formants widening.

On the other hand, the spectrogram of SSFV modification (Fig. 10 (e)) has been able to highlight the speech formants in the presence of low frequency noise spectrum better than NM, ER and NU-TSM modifications. This is due to the shifting of low frequency speech formants upward away from the region of noise spectrum. Additionally, it can also be observed from FTSFE spectrogram shown in Fig. 10 (f) that the low frequency spectrum is preserved better than any other modifications with higher visibility of formants. This is because FTSFE utilizes the combined properties of individual modifications. Hence, it can be concluded that the FTSFE has been an adequate compromise between NU-TSM and SSFV.

7. Conclusion and future scope

In this work, we focused on improving speech intelligibility for in-car applications by transforming normal speech to Lombard-like speech. This transformation is achieved by using a set of voice conversion effects namely, non uniform-time scale modification, smoothed shifting of formants for voiced segments and energy redistribution between voiced and unvoiced segments. The subjective and objective evaluations have shown significant voice intelligibility improvement for normal hearing and hearing impaired listeners using the proposed modifications. Additionally, the combined model gathers the greatest number of participants with positive relative threshold in comparison to other modifications. The spectrographic evaluation for the hybrid model also highlights the visibility of speech formants in the noise better than any other modifications.

Future scope would be to investigate new methods which will consider hearing abilities of impaired persons in shifting the formants away from the region of their loss. Additionally, it would be interesting to investigate the effect of pitch shifting or pitch modulation for intelligibility improvement.

Acknowledgement

This work is supported by the French National Research Agency (ANR) as a part of AIDA Project Edition 2013.

References

Amano-Kusumoto, A., Hosom, J.-P., 2011. A review of research on speech intelligibility and correlations with acoustic features. Center for Spoken Language Understanding, Oregon Health and Science University (Technical Report CSLU-011-001).
Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A., Kitamura, T., 2002. Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments. Acoust. Sci. Technol. 23 (4), 229–232.
Barker, J., Cooke, M., 2007. Modelling speaker intelligibility in noise. Speech Commun. 49 (5), 402–417.
Bond, Z.S., Moore, T.J., 1994. A note on the acoustic-phonetic characteristics of inadvertently clear speech. Speech Commun. 14 (4), 325–337.
Brown, R.G., 1959. Statistical Forecasting for Inventory Control. McGraw/Hill.
CND, 2015. National college of audioprothesist - speech audiometry compact disk (text in french).
Cooke, M., Mayo, C., Valentini-Botinhao, C., 2013. Intelligibility-enhancing speech modifications: the hurricane challenge.. In: Interspeech, pp. 3552–3556.
Desai, S., Raghavendra, E.V., Yegnanarayana, B., Black, A.W., Prahallad, K., 2009. Voice conversion using artificial neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3893–3896.
Erogul, O., Karagoz, I., 1998. Time-scale modification of speech signals for language-learning impaired children. In: 2nd IEEE International Conference on Biomedical Engineering Days. IEEE, pp. 33–35.
Ferguson, S.H., Kewley-Port, D., 2002. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 112 (1), 259–271.
Garnier, M., Henrich, N., 2014. Speaking in noise: how does the lombard effect improve acoustic contrasts between speech and ambient noise? Comput. Speech Lang. 28 (2), 580–597.
Hagmüller, M., Kubin, G., 2006. Poincaré pitch marks. Speech Commun. 48 (12), 1650–1665.
Hazan, V., Markham, D., 2004. Acoustic-phonetic correlates of talker intelligibility for adults and children. J. Acoust. Soc. Am. 116 (5), 3108–3118.

Hazan, V., Simpson, A., 1998. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. Speech Commun. 24 (3), 211–226.
Hodoshima, N., Arai, T., Kusumoto, A., 2002. Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments. In: Proc. Forum Acusticum, Sevilla.
Hu, Y., Loizou, P.C., 2007. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Am. 122 (3), 1777–1786.
Junqua, J.-C., 1993. The lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93 (1), 510–524.
Kawahara, H., 1997. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2. IEEE, pp. 1303–1306.
Kim, G., Loizou, P.C., 2010. Improving speech intelligibility in noise using environment-optimized algorithms. IEEE Trans. Audio Speech Lang. Process. 18 (8), 2080–2090.
Kotnik, B., Höge, H., Kacic, Z., 2006. Evaluation of pitch detection algorithms in adverse conditions. In: Proc. of the Speech Prosody. Citeseer.
Kupryjanow, A., Czyzewski, A., 2012. Methods of improving speech intelligibility for listeners with hearing resolution déficit. Diagn. Pathol. 7 (1), 1–18.
Laures, J.S., Bunton, K., 2003. Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. J. Commun. Disord. 36 (6), 449–464.
Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC press.
Loizou, P.C., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Audio Speech. Lang. Process. 19 (1), 47–56.
Lombard, E., 1911. Le signe de l'elevation de la voix. Ann. Maladies Oreille, Larynx, Nez, Pharynx 37 (101–119), 25.
Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. J. Acoust. Soc. Am. 124 (5), 3261–3275.
Lu, Y., Cooke, M., 2009. The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Commun. 51 (12), 1253–1262.
Ma, J., Hu, Y., Loizou, P.C., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125 (5), 3387–3405.
Machado, A.F., Queiroz, M., 2010. Voice conversion: a critical survey. Proc. Sound Music Comput. (SMC) 1–8.
Moon, S.-J., Lindblom, B., 1994. Interaction between duration, context, and speaking style in english stressed vowels. J. Acoust. Soc. Am. 96 (1), 40–55.
Moulines, E., Laroche, J., 1995. Non-parametric techniques for pitch-scale and time-scale modification of speech. Speech Commun. 16 (2), 175–205.
Nathwani, K., Daniel, M., Richard, G., David, B., Roussarie, V., 2016. Formant shifting for speech intelligibility improvement in car noise environment. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5375–5379.
Nilsson, M., Soli, S.D., Sullivan, J.A., 1994. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Am. 95 (2), 1085–1099.
Nurminen, J., Popa, V., Tian, J., Tang, Y., Kiss, I., 2006. A parametric approach for voice conversion. In: TCSTAR WSST, pp. 225–229.
Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice Hall.
Rao, K.S., Yegnanarayana, B., 2006. Voice conversion by prosody and vocal tract modification. In: 9th IEEE International Conference on Information Technology (ICIT). IEEE, pp. 111–116.
Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 749–752.
Skowronski, M.D., Harris, J.G., 2006. Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments. Speech Commun. 48 (5), 549–558.
Steeneken, H.J., Hansen, J.H., 1999. Speech under stress conditions: overview of the effect on speech production and on system performance. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2079–2082.
Stylianou, Y., 2009. Voice transformation: a survey. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3585–3588.
Taal, C.H., Hendriks, R.C., Heusdens, R., 2014. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. Comput. Speech Lang. 28 (4), 858–872.
Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4214–4217.
Taal, C.H., Jensen, J., 2013. Sii-based speech preprocessing for intelligibility improvement in noise. In: INTERSPEECH, pp. 3582–3586.
Taghia, J., Martin, R., 2014. Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (1), 6–16.
Taghia, J., Martin, R., Hendriks, R.C., 2012. On mutual information as a measure of speech intelligibility. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 65–68.
Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., Soli, S.D., Giguère, C., 2005. Adaptation of the hint (hearing in noise test) for adult canadian francophone populations: adaptación del hint (prueba de audición en ruido) para poblaciones de adultos canadienses francófonos. Int. J. Audiol. 44 (6), 358–361.
Van Summers, W., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A., 1988. Effects of noise on speech production: Acoustic and perceptual analyses. J. Acoust. Soc. Am. 84 (3), 917–928.
Vincent, E., Bertin, N., Badeau, R., 2010. Adaptive harmonic spectral decomposition for multiple pitch estimation. IEEE Trans. Audio Speech Lang. Process. 18 (3), 528–537.
Yang, H., Guo, W., Liang, Q., 2008. A speaking rate adjustable digital speech repeater for listening comprehension in second-language learning. In: IEEE International Conference on Computer Science and Software Engineering, 5. IEEE, pp. 893–896.
Zahorian, S.A., Hu, H., 2008. A spectral/temporal method for robust fundamental frequency tracking. J. Acoust. Soc. Am. 123 (6), 4559–4571.
Zhang, M., Tao, J., Tian, J., Wang, X., 2008. Text-independent voice conversion based on state mapped codebook. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4605–4608.