On automatic recognition of spectrally reduced speech synthesized from amplitude and frequency modulations

Cong Thanh Do, Dominique Pastor, André Goalic

Collection des rapports de recherche
de TELECOM Bretagne

RR-2009001-SC

On Automatic Recognition of Spectrally Reduced Speech Synthesized From Amplitude and Frequency Modulations

Cong-Thanh Do (TELECOM Bretagne, Lab-STICC)
Dominique Pastor (TELECOM Bretagne, Lab-STICC)
André Goalic (TELECOM Bretagne, Lab-STICC)

On Automatic Recognition of Spectrally Reduced Speech Synthesized From Amplitude and Frequency Modulations

Sur la Reconnaissance Automatique de Parole Réduite Spectralement Synthétisée à partir des Modulations d'Amplitude et de Fréquence

Rapport Interne Institut TELECOM / TELECOM Bretagne

Cong-Thanh Do, Dominique Pastor, André Goalic

Institut TELECOM; TELECOM Bretagne,
Lab-STICC, CNRS UMR 3192,
Technopôle de Brest Iroise, CS 83818
29238 BREST Cedex 3,
FRANCE
{thanh.do, dominique.pastor, andre.goalic}@telecom-bretagne.eu

Abstract

This report investigates the behavior of automatic speech recognition (ASR) system with spectrally reduced speech (SRS) synthesized from subband amplitude modulations (AMs) and frequency modulations (FMs). Acoustic analysis shows that the resynthesis of SRS from only AM components helps alleviate certain non-linguistic variabilities in the original speech signal. When the SRS spectral resolution is sufficiently good, this alleviation not only has no consequence but also yields comparable or even better ASR word accuracy compared to that attained with original clean speech signal. In contrast, FM components support human speech recognition but yield no significant improvement in terms of ASR word accuracy when the SRS spectral resolution is sufficiently good.

Keywords

Automatic speech recognition, spectrally reduced speech, amplitude modulation, frequency modulation.

Résumé

Ce rapport étudie le comportement de système de reconnaissance automatique de la parole (ASR) avec la parole réduite spectralement (SRS) synthétisée à partir des modulations d'amplitude (AM) et de fréquence (FM). Les analyses d'acoustique montrent que la resynthèse de la SRS à partir des composants AM permet à s'affranchir certaines variabilités non-linguistiques dans le signal de parole original. Lorsque la résolution spectrale de la SRS est suffisamment bonne, cette simplification n'a aucune conséquence tout en gardant le précision de mot (word accuracy) de l'ASR comparable ou meilleure que celui atteinte par le signal de parole propre original. En revanche, les composants FM supportent la reconnaissance de la parole de l'être humain, mais n'apportent pas d'amélioration significative en terme de précision de mot de l'ASR lorsque la résolution spectrale de la SRS est suffisamment bonne.

Mots-clés

Reconnaissance automatique de la parole, parole réduite spectralement, modulation d'amplitude, modulation de fréquence.

Contents

1 Introduction

Speech communication among humans is stable despite the great variability in the speech signal. There are many sources of acoustic variance in the speech signal that are not directly associated with the linguistic message, including (1) acoustic degradations (e.g., constant or slowly varying additive noise, microphone frequency response, talker or microphone movement, etc.) and (2) speech production variations (e.g., accent and dialect, speaking style, acoustic variability due to specific states of health and mental state, etc.) [1].

The spectrum of sound is considered to be an important correlate of phonetic quality of speech sounds [2]. However, the prime carrier of the linguistic information are changes of the spectral envelopes of the speech signal [3]. Indeed, the relatively narrow bandwidth of the spectral envelope modulations is created by the relatively slow motion of the vocal tract during speech production. In this view, the "message" (the signal containing information of vocal tract motions with a narrow bandwidth) modulates the "carrier" signal (high frequency) analogous to the amplitude modulation (AM) used in radio communications. The major linguistically significant information in speech is contained in the details of this low frequency vocal tract motion (i.e., the spectral envelope of the speech signal) [4].

In an automatic speech recognition (ASR) system, the speech analysis module performs feature extraction using signal processing techniques to compensate for variability in speaker characteristics as well as acoustic environment. It can help alleviate non-linguistic components of speech signal and improve reliability of ASR in realistic environments [2]. Most state-of-the-art ASR systems use Mel cepstrum [5] or Perceptual Linear Predictive (PLP) analysis [6] of speech in the speech analysis module. In the calculation of the feature vectors, the speech spectrum at a given time instant is derived from a segment of speech which is short enough (of the order of 10-20 ms) to be assumed to result from a stationary process. The phase of such a short-term spectrum is typically discarded. The fine structure of this short-term spectrum (which carries information about the voice source) is most often discarded as well, and it is the spectral envelope of the speech short-term spectrum which provides a starting point for most speech features used in ASR [2].

So what will happen if certain speech non-linguistic variabilities can be reduced before the resulting signal being fed to the speech analysis module of an ASR system? Will this reduction have any influence on the performance of an ASR system and if so, how can we adjust the degree of this influence? We thus resynthesize a new kind of speech, called spectrally reduced speech (SRS), from original clean speech and then use this SRS in an ASR system

(Fig. 1). By resynthesizing SRS from original clean speech signal, certain types of speech non-linguistic variabilities can be alleviated. Shannon et al.,

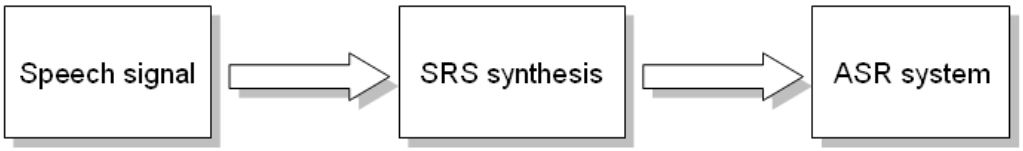

Figure 1: Using SRS in ASR. SRS is first synthesized from original speech signal and then is used in ASR.

[7] synthesized SRS by using amplitude modulations (AMs) extracted from 3 or 4 frequency bands of speech to modulate white noise and then summed up these subband signals to obtain SRS. Despite the great reduction of spectral information, perceptive tests have shown that this kind of speech could support human speech recognition in quiet environment. Recently, Nie et al., [8] have suggested that the slowly varying frequency modulation (FM) should be integrated in the SRS described in [7] to support human speech recognition in noisy environment. Indeed, acoustic analysis have shown that slowly varying FM components preserve dynamic information regarding speech formant transitions and fundamental frequency movements; these may help human recognize speech in noise [8]. In this letter, we will use the SRSs proposed in [7] and [8] for testing with ASR because these SRSs make it possible to reduce certain non-linguistic variabilities in the resynthesized speech signal. Above a certain SRS spectral resolution, we will see that the ASR word accuracy [9] attained with SRS synthesized from only AM components, is surprisingly comparable to or even better than that attained with the original clean speech signal. This result suggests many prospects for future applications.

2 SRS Synthesis Using AM-FM Model of Speech

A speech signal $s(t)$ can be approximated by a sum of N band-limited components, $s_k(t)$, $k = 1, \cdots, N$.

$$s(t) \cong \sum_{k=1}^{N} s_k(t) \tag{1}$$

Each subband component $s_k(t)$ can be considered as an AM-FM signal [10]

which contains both amplitude and frequency modulations

$$s_k(t) = m_k(t)\cos\left[2\pi f_{ck}t + 2\pi\int_0^t g_k(\tau)d\tau + \theta_k\right] \tag{2}$$

where the signals $m_k(t)$ and $g_k(t)$ are the amplitude modulation and the frequency modulation in the k^{th} subband whereas f_{ck} and θ_k are the k^{th} subband central frequency and the initial phase, respectively. In this paper, we use the term AM-only SRS to designate SRS synthesized by using only the AM components in the subbands of the decomposed speech signal. To synthesize AM-only SRS, the speech signal is first decomposed into several frequency bands and AMs are extracted from the subband signals by full wave rectification followed by lowpass filtering. The extracted AMs of the subband signals are then used to modulate either white noise or sinusoids whose frequencies equal the central frequencies of the frequency subbands [7]. The modulated subband signals from all the frequency bands are then summed up to construct the AM-only SRS. The mathematical expression $\widehat{s}_{\mathrm{AM}}(t)$ of AM-only SRS can be written as follows

$$\widehat{s}_{\mathrm{AM}}(t) = \sum_{k=1}^{N}\widetilde{m_k}(t)\cos(2\pi f_{ck}t) \tag{3}$$

where $\widetilde{m_k}(t)$ is the lowpass filtered AM in the k^{th} subband.

Besides, we use the acronym AM+FM SRS to designate SRS synthesized from both AM and FM components. The frequency amplitude modulation encoding (FAME) strategy [8], which is used to synthesize AM+FM SRS, involves two steps, analysis and synthesis (Fig. 2). In the analysis stage, the speech signal is first decomposed into several frequency bands by using a critical filter bank having logarithmically increasing central frequencies to mimic the human cochlear filters. Then, two independent parallel pathways are employed to extract the slowly varying AM and FM cues from each of the filtered subband signal. In the FM extraction pathway, the extracted FM is band-limited and further lowpass filtered to obtain the slowly varying FM. In the synthesis, the slowly varying FM signal is added to the original central frequency f_{ck}, then integrated to recover the original phase information and, finally, multiplied by the slowly varying AM signal $\widetilde{m_k}(t)$ extracted from the same subband to recover the original subband signal [8]. Finally, by summing the recovered subband signals from all the subbands, AM+FM SRS is obtained. This signal, $\widehat{s}_{\mathrm{AM+FM}}(t)$, can be written as follows

$$\widehat{s}_{\mathrm{AM+FM}}(t) = \sum_{k=1}^{N}\widetilde{m_k}(t)\cos\left[2\pi\int_0^t(\widetilde{g_k}(\tau) + f_{ck})d\tau\right] \tag{4}$$

where $\widetilde{g}_k(t)$ stands for the slowly varying FM in the k^{th} subband.

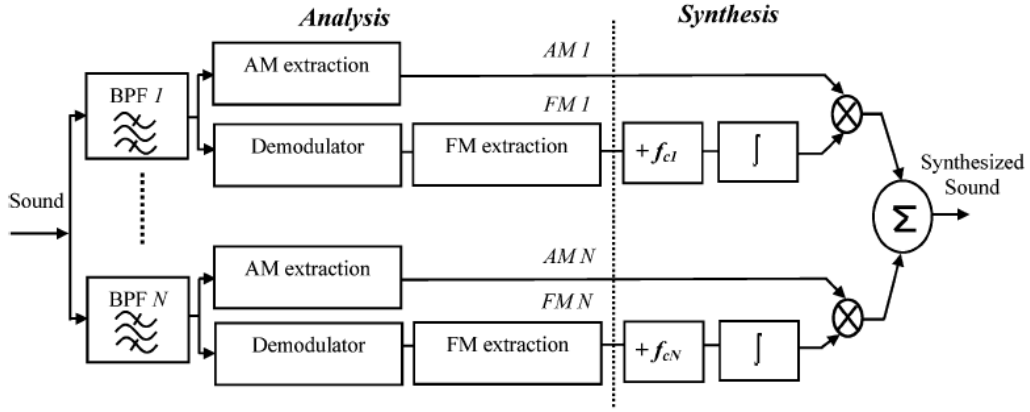

Figure 2: Functional block diagram of speech processing in the FAME strategy for AM+FM SRS synthesis [8].

Detailing further the FM extraction algorithm of FAME in Fig. 3, we can see that this is a conventional FM extraction diagram followed by two blocks "Band limit" and "LPF 3" to control the FM bandwidth and the FM rate, respectively. The subband signal $s_k(t)$ is first multiplied with the outcome of a quadrature oscillator with central frequency f_{ck}. This manipulation is equivalent to shifting the spectrum of $s_k(t)$ to 0 and $2f_{ck}$ in the frequency domain. After the lowpass filterings performed by *LPF 2* and *LPF 2'*, the in-phase signal, a, and the out-of-phase signal, b, of the subband signal $s_k(t)$, are extracted. In the discrete implementation, the FM signal is calculated by the following formula [8]

$$\text{FM} = \frac{b\Delta a - a\Delta b}{2\pi(a^2 + b^2) \times T_s} \tag{5}$$

where Δ is the differentiation and T_s is the sampling period. The mathematical development of the FM extraction process can be found in [8]. "Band limit" is used to avoid undesirable cross-talk between adjacent bands whereas *LPF 3* is used to set the FM rate lower than 400 Hz so that it may be perceived by human cochlear-implant listeners [8]. Using the slowly varying AM signal $\widetilde{m_k}(t)$ to modulate the FM full frequency range (FR) signal $g_k(t)$ taken at **c** (Fig. 3), we synthesize the AM+FM SRS with full frequency range FM component, termed FR AM+FM SRS. In the same way, by using the slowly varying FM signal $\widetilde{g}_k(t)$ taken at **d**, we can synthesize the AM+FM SRS with FM slowly varying (SV) component, termed SV AM+FM SRS.

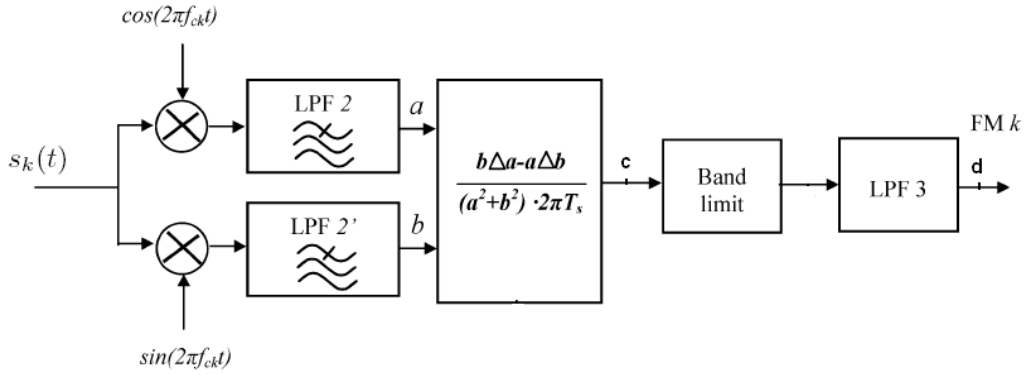

Figure 3: Algorithm for extracting FM components in the k^{th} subband of the FAME strategy [8].

The analysis filter bank in this work contains bandpass filters whose central frequencies and bandwidths match those of the critical bands [11]. Each bandpass filter is constructed using a fourth-order elliptic bandpass filter having a stop-band attenuation of 50 dB and a 2 dB ripple in the pass-band to be consistent with [8]. The original speech signal $s(t)$ is thus decomposed into subband signals $s_k(t),\ k = 1,\ \cdots,\ N$ with N taking values in $\{4, 6, 8, 10, 16, 24, 32\}$. Subsequently, fourth-order elliptic lowpass filter with 50 Hz cutoff frequency is used in the AM extraction. In the FM extraction, the two filters used to determine the FM depth or the FM bandwidth, *LPF 2* and *LPF 2'*, are also fourth-order lowpass Bessel filters. Their cutoff frequencies are set to either 500 Hz or the analysis subband filter bandwidth whenever this bandwidth is less than 500 Hz [8]. Finally, the cutoff frequency of the FM-rate filter, *LPF 3*, is set to 400 Hz. After AM and FM extractions, the corresponding AM-only SRS, FR AM+FM SRS and SV AM+FM SRS are synthesized. Fig. 4 shows the frequency response of an analysis filter bank containing 16 fourth-order elliptic bandpass filters.

3 Acoustic Analysis of Synthesized SRS

Fig. 5 shows spectrograms of a continuous speech utterance "one oh four six", pronounced by a female speaker, from the TIDIGITS speech database. The spectrogram of the original clean speech utterance is shown in 5(a) whereas 5(b), 5(c), and 5(d) show the spectrograms of the 4-subband synthesized SRSs; 5(b): 4-subband AM-only SRS, 5(c): 4-subband SV AM+FM SRS, and 5(d): 4-subband FR AM+FM SRS. Similarly, Fig. 6 presents the

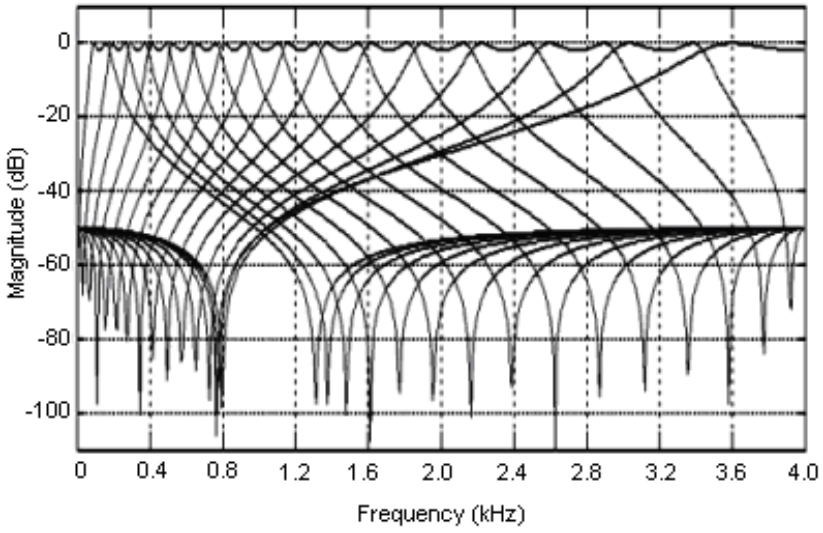

Figure 4: Frequency response of an analysis filter bank containing 16 fourth-order elliptic bandpass filters used for speech signal decomposition. The speech signal is sampled at 8 kHz.

spectrograms of 16-subband synthesized SRSs for the same utterance; 6(a): original clean speech, 6(b): 16-subband AM-only SRS, 6(c): 16-subband SV AM+FM SRS, and 6(d): 16-subband FR AM+FM SRS. We can see that the formant transitions (e.g., at positions of the arrows in Fig. 5 and Fig. 6) are present in 5(a), 5(c), and 5(d) (idem for 6(a), 6(c), and 6(d)) but not in 5(b). Fig. 5(b) is the spectrogram of the 4-subband AM-only SRS in which the FM component is not integrated. However, it seems that the formant transitions are still preserved in the 16-subband AM-only SRS (see Fig. 6(b)), in which the spectral resolution is sufficiently good.

The blue line in each spectrogram, estimated by using Praat software [12], represents the speech fundamental frequency (f_0) as a function of time. The range of the fundamental frequency is figured on the right-hand side vertical axis of each spectrogram. Using the extracted f_0 contour of the original clean speech as the reference, we can remark that the f_0 information is not correctly estimated from the AM-only SRSs, whether we use 4 or 16 subbands. Another remark concerns the f_0 information in the AM+FM SRSs. The f_0 information is not correctly estimated in the 4-subband SV AM+FM SRS but its estimation in the 4-subband FR AM+FM SRS is sufficiently good. This does not happen with the 16-subband SV AM+FM SRS (6(c)) and the 16-subband FR AM+FM SRS (6(d)) where Praat estimates correctly f_0. We can therefore conclude that FM components, especially the

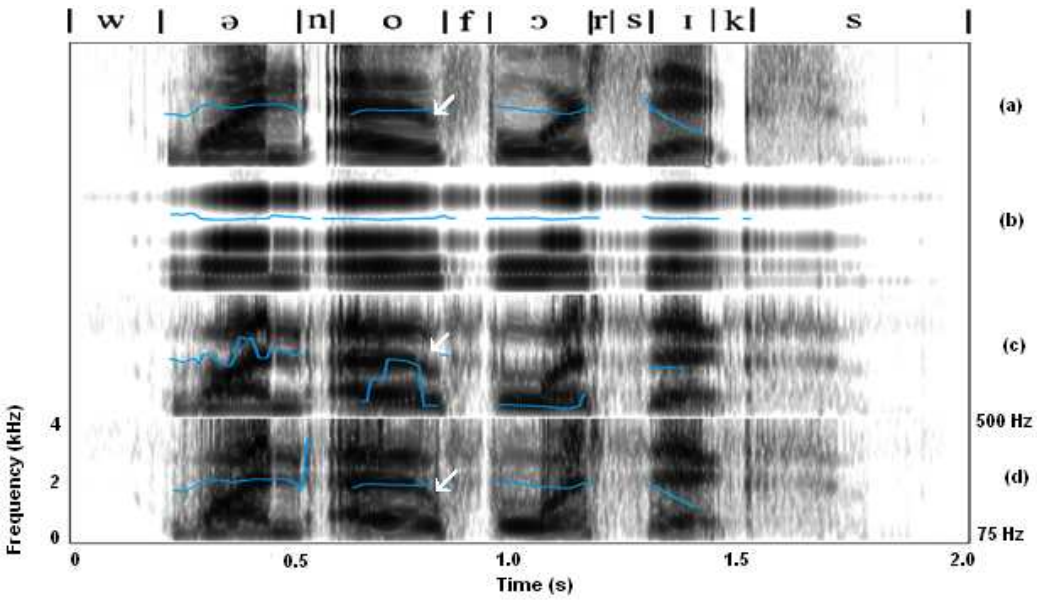

Figure 5: Spectrograms of synthesized SRSs of the speech utterance "one oh four six" from the TIDIGITS database, pronounced by a female speaker. (a) original clean speech, (b) 4-subband AM-only SRS, (c) 4-subband SV AM+FM SRS, (d) 4-subband FR AM+FM SRS. Arrows indicate formant transitions from vowel /o/ in "oh" to consonant /f/ in "four". These formant transitions are absent in (b), the spectrogram of the 4-subband AM-only SRS. The blue lines, estimated by Praat [12], represent the speech fundamental frequency f_0. The f_0 frequency range is [75 Hz - 500 Hz] (see the right-hand side vertical axis).

rapidly varying ones, contain information regarding the speech fundamental frequency which is a speaker-dependent information. However, even in case of missing FM rapidly varying components, increasing the SV AM+FM SRS spectral resolution makes it possible to retrieve the f_0 information.

4 Testing SRS with an ASR System

We construct a speaker-independent ASR system for testing the behavior of ASR with SRSs. The TIDIGITS, which is a small vocabulary, continuous speech database and the HTK speech recognition software [9] are used for the training of the phoneme acoustic models. Each phoneme is modeled by a context-dependent three-state left-to-right triphone Hidden Markov Model

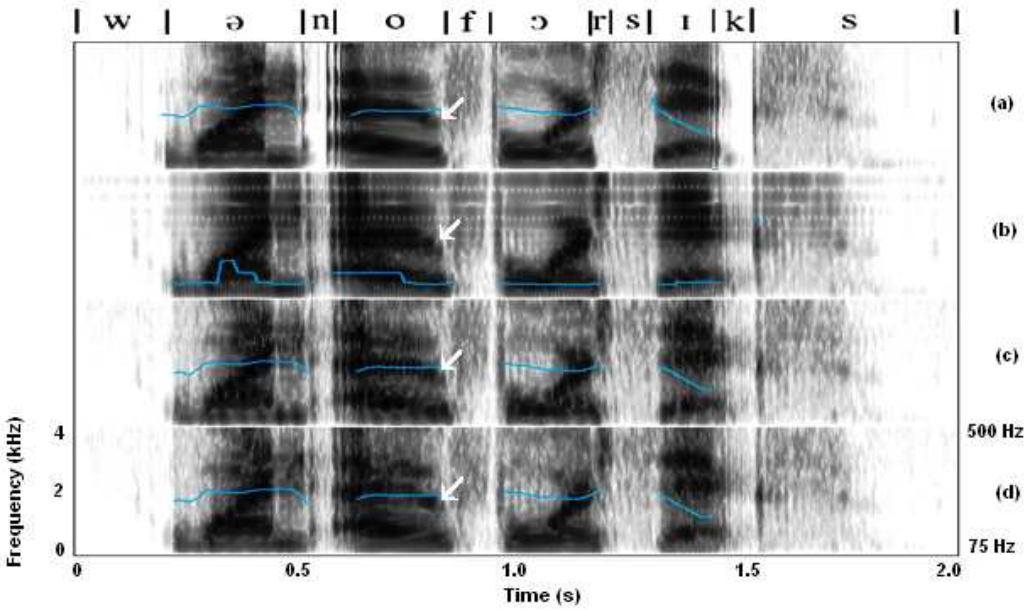

Figure 6: Spectrograms of the synthesized SRSs of the same utterance as in Fig. 5. (a) original clean speech, (b) 16-subband AM-only SRS, (c) 16-subband SV AM+FM SRS, (d) 16-subband FR AM+FM SRS.

(HMM). The output distributions are modeled by multivariate Gaussian Mixtures Densities. Each mixture consists of 16 multivariate Gaussian components. Feature vectors consist of 13 Mel Frequency Cepstral Coefficients (MFCCs) in which the first coefficient C_0 is used as the energy component [9]. The MFCCs are calculated from every Hamming windowed speech frame of 25 ms length. The overlap between two adjacent frames is 15 ms. The delta and acceleration coefficients [9] are appended to the MFCCs to provide 39-dimensional feature vectors. 250 utterances from the TIDIGITS clean speech test database are selected for the tests. From this set of original clean speech signals, we synthesize the AM-only SRSs, the SV AM+FM SRSs, and the FR AM+FM SRSs. Table 1 shows the ASR word accuracies of synthesized SRSs by using the triphone HMMs trained on the TIDIGITS clean speech training database.

The recognition results show that among the three kinds of SRS studied, the AM-only SRSs give the best ASR word accuracies when the SRSs are synthesized on the basis of AM and FM components of at least 16 subbands. We thus use the model of 16-subband AM-only SRS, the AM-only SRS who gave the best ASR word accuracy in Table 1, to resynthesize the TIDIGITS

Table 1: ASR word accuracies (in %) of synthesized SRSs using triphone HMMs trained on the TIDIGITS clean speech training database. The ASR word accuracy computed on 250 original clean speech utterances is 99.76%.

SRS	Number of frequency bands						
	4	6	8	10	16	24	32
AM-only	35.57	41.05	84.23	98.48	**99.82**	**99.70**	**99.57**
SV AM+FM	64.43	**97.87**	**98.42**	**99.09**	99.33	98.36	96.41
FR AM+FM	**82.95**	92.81	93.61	94.15	94.76	94.88	94.70

training database. Using the triphone HMMs trained on this new training database for recognition of the testing SRSs, we obtain the ASR word accuracies of Table 2. These results confirm the conclusion mentioned above:

Table 2: ASR word accuracies (in %) of synthesized SRSs using triphone HMMs trained on the TIDIGITS training database resythesized from the 16-subband AM-only SRS model. The ASR word accuracy computed on the same 250 original clean speech utterances mentioned above is 99.45%.

SRS	Number of frequency bands						
	4	6	8	10	16	24	32
AM-only	34.59	39.77	89.89	**99.39**	**99.76**	**99.70**	**99.70**
SV AM+FM	40.13	85.08	**95.43**	97.99	98.96	94.58	87.76
FR AM+FM	**68.94**	**88.86**	91.60	91.72	94.28	94.40	93.85

AM-only SRSs give the best ASR word accuracies when the SRSs spectral resolution is sufficiently good (10 subbands in this case).

5 Conclusions

We have investigated the behavior of ASR with SRSs synthesized from subband AMs and FMs. For the small vocabulary, continuous speech database (TIDIGITS) studied in this letter, the following conclusions can be formulated:

- Above a certain SRS spectral resolution, AM-only SRS gives the best ASR word accuracy among the three kinds of SRS studied.
- We can use AM-only SRSs synthesized from 10 subbands or more to achieve comparable or even better ASR word accuracies compared to those attained with original clean speech (see Table 1 and Table 2).
- FM components support human speech recognition [8] but yield no significant improvement in terms of ASR word accuracy when the SRSs spectral resolution is sufficiently good.

The acoustic analysis of synthesized SRSs has shown that even though the formant transitions are not clearly visible and the fundamental frequency information is absent in the AM-only SRS, this kind of SRS still provides the best ASR word accuracy when the SRS spectral resolution is sufficiently good. This suggests that certain non-linguistic variabilities in the speech signal can actually be alleviated via SRS resynthesis. Using appropriate synthesized SRSs in ASR can help achieve comparable or even better ASR word accuracies compared to those attained with original clean speech. Future work will extend the study to large vocabulary and focus on finding new kinds of SRS that make it possible to reduce more speech acoustic variances while yielding good ASR word accuracy.

6 Acknowledgement

The authors would like to acknowledge M.Vijay Kumar for his assistance during the summer 2008.

References

[1] N. Morgan, H. Bourlard, and H. Hermansky, Speech processing in the auditory system. Springer, 2004, chapter. Automatic speech recognition: an auditory perspective, pp. 309338.

[2] H. Hermansky, Mel cepstrum, deltas, double-deltas,.. what else is new? Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, vol. 1, pp. 866871, 1999.

[3] H. Hermansky, Should recognizers have ears? Speech Communication, vol. 25, no. 1-3, pp. 327, Aug. 1998.

[4] C. Avendano, L. Deng, H. Hermansky, and B. Gold, Speech processing in the auditory system. Springer, 2004, ch. The analysis and representation of speech, pp. 63100.

[5] S. B. Davis and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences, IEEE Trans. ASSP, vol. 28, no. 4, pp. 357366, Aug. 1980.

[6] H. Hermansky, Perceptual linear predictive (plp) analysis of speech, J. Acoust. Soc. Am., vol. 87, no. 4, pp. 17381752, Apr. 1990.

[7] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, Speech recognition with primarily temporal cues, Science, vol. 270, no. 5234, 1995.

[8] K. Nie, G. S. Stickney, and F. G. Zeng, Encoding frequency modulation to improve cochlear implant performance in noise, IEEE Trans. Biomed. Eng., vol. 52, no. 1, pp. 6473, Jan. 2005.

[9] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK book (for HTK version 3.2.1). Cambridge, UK: Cambridge University Engineering Department, 2002.

[10] P. Maragos, J. F. Kaiser, and T. F. Quatieri, Energy separation in signal modulations with application to speech analysis, IEEE Trans. Sig. Process., vol. 41, no. 10, pp. 30243051, Oct. 1993.

[11] T. S. Gunawan and E. Ambikairajah, Speech enhancement using temporal masking and fractional bark gammatone filters, in Proc. of the 10th Australian Internatl. Conf. on Speech Sci. & Tech., December 8 - 10, Sydney, Australia, Dec. 2004, pp. 420425.

[12] P. Boersma, Praat, a system for doing phonetic by computer, Glot International, pp. 341345, 2001.

www.telecom-bretagne.eu

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 3
France
Tél. : + 33 (0)2 29 00 11 11
Fax : + 33 (0)2 29 00 10 00

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
France
Tél. : + 33 (0)2 99 12 70 00
Fax : + 33 (0)2 99 12 70 19

Campus de Toulouse
10, avenue Edouard Belin
BP 44004
31028 Toulouse Cedex 04
France
Tél. : +33 (0)5 61 33 83 65
Fax : +33 (0)5 61 33 83 75

Imprimé à TELECOM Bretagne
Dépôt légal : janvier 2009
ISSN : 1255-2275

TELECOM
Bretagne