(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 2 372 707 A1

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
05.10.2011 Bulletin 2011/40

(51) Int Cl.:
G10L 21/02 $^{(2006.01)}$

(21) Application number: **10156530.7**

(22) Date of filing: **15.03.2010**

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO SE SI SK SM TR
Designated Extension States:
AL BA ME RS

(71) Applicant: **Svox AG**
8048 Zürich (CH)

(72) Inventors:
- **Withopf, Jochen**
  97941 Tauberbischofsheim (DE)
- **Hannon, Patrick**
  89077 Ulm (DE)
- **Krini, Mohamed**
  89075 Ulm (DE)
- **Schmidt, Gerhard Uwe**
  24248 Mönkeberg (DE)

(74) Representative: **Stocker, Kurt**
Büchel, von Révy & Partner
Zedernpark
Bronschhoferstrasse 31
9500 Wil (CH)

(54) Adaptive spectral transformation for acoustic speech signals

(57) The present invention relates to a method for adaptive transformation of the frequency spectrum of a windowed speech signal. According to the present invention, frequency compression is achieved by applying frequency compression functions, which are dependent upon characteristics of the current frame. Formant boosting takes place using formant-dependent functions to increase the contrast between the formants and the non-formant frequencies in the frequency spectrum of the current frame. Generally, frequencies below a defined threshold are compressed linearly, corresponding to no compression when the slope is equal to 1, while the compression rate of higher frequencies varies over time and frequency depending on the current frame features. Thus features of the input signal located above the frequency threshold are moved into the low-pass range and are audible in the transmitted output signal. The feature based processing is designed to enhance the understandability of speech in the transmitted bandwidth and to increase recognition rates in an automatic speech recognition module.

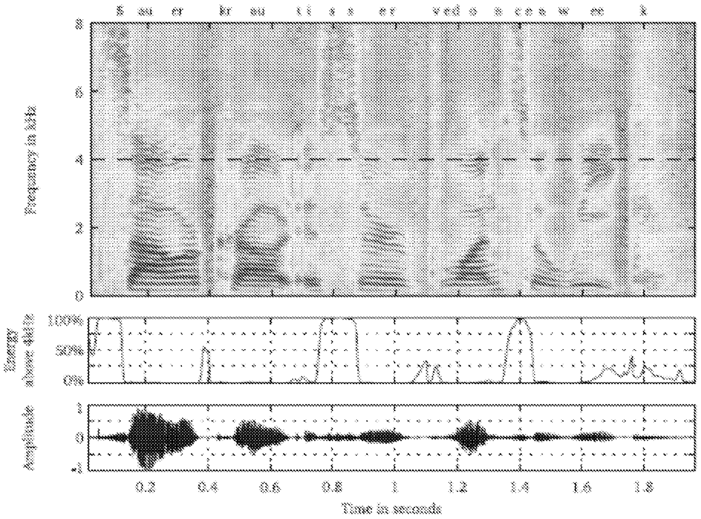


Fig. 1

Printed by Jouve, 75001 PARIS (FR)

Description

Technical Field

[0001] The present invention generally relates to speech synthesis technology.

Background of the invention

[0002] In most telecommunication systems, speech signals are not transmitted with their full analog bandwidth in order to use the available transmission channel more efficiently. For telephone communication, the signal bandwidth is typically limited to less than 4 kHz, even though there are signal components up to 8 kHz and higher in the original speech signal. This band limitation has little or no effect on the intelligibility of voiced sounds, but fricatives such as /s/, /sh/, /ch/, /z/ or /f/ may be lost completely. In Fig. 1 a spectrogram of the utterance "Sauerkraut is served" is depicted. The phoneme /s/ is spoken at approximately 0, 0.8, and 1.4 seconds. At these times almost 100% of the signal energy is above 4 kHz. With a bandwidth limited to less than 4 kHz the speech intelligibility may still be sufficient because the listener is often able to predict missing phonemes from syntax and the context of what is being said. However, errors arise if such a prediction is not possible, e.g., because names or unknown words of a foreign language are transmitted. Furthermore, the phoneme /s/ is important in the English language to indicate the plural and possessive pronouns.

[0003] Various spectral compression schemes exist that aim to present signals from one frequency range in another more useful range of the frequency spectrum. The side-effects of the schemes vary according to the limitations of the candidate identification techniques, artifacts resulting from the signal processing, as well as high computational complexity and delay in the signal analysis stage.

[0004] P. Patrick, R. Steele, and C. Xydeas, Frequency compression of 7.6 kHz speech into 3.3 kHz bandwidth, 31 (5):692-701, May 1983 describes a system for retaining information under frequency compression. This system consists of frequency mapping at the transmitter and demapping at the receiver side. According to a frequency compression factor c, every cth sample is retained in the compressed magnitude spectrum, so that it occupies only the frequency range that is available for transmission. At the receiver side, the received frequency components are spaced out to their correct locations and the magnitude of the missing components is found by linear interpolation. The phase is chosen randomly.

[0005] The compression factor c can also be set frequency dependently to give higher or lower compression in certain regions. Three different mapping laws, which are tailored for certain phonemes, and one that simply corresponds to a band limitation, are used in the system. For switching between these laws, the signal is first classified as voiced or unvoiced speech based on its autocorrelation. Mapping is only applied for unvoiced speech. Then the mapping and demapping procedure is applied according to the same laws.

[0006] It is important to note that this system needs to transmit side information about the mapping laws that have been used to ensure correct demapping. Additionally, the receiver must be able to handle this side information and to perform the demapping.

[0007] D. A. Heide and G. S. Kang, Speech enhancement for bandlimited speech. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, volume 1, pages 393-396, May 12-15, 1998 describes a method that works without transmitting any side information. A classification of the speech sound is based on the spectral centroid of the spectrum, f_c. For voiced speech, $f_c \approx 2$ kHz usually holds.

[0008] When the spectral centroid is greater, portions of the spectrum above 4 kHz are added into the 3 - 4 kHz bands according to the following rules:

fc = 2 kHz: No spectral translation is performed.

fc > 3 kHz: The 4 - 5 kHz band is added into the 3 - 4 kHz band.

fc > 4 kHz: The 5 - 6 kHz band is also added into the 3 - 4 kHz band.

fc > 5 kHz: The 6 - 7 kHz band is also added into the 3 - 4 kHz band.

[0009] Uncontrolled artifacts are resulting from this signal processing. There is no general improvement of the speech intelligibility.

[0010] Another application area of frequency mapping methods is hearing aids. People with hearing impairment usually suffer from a bad perception of high frequency sounds. The traditional approach is to strongly amplify these critical frequency regions. However, for some people, hearing sensitivity is so poor at high frequencies that sufficient gain for achieving audibility cannot be provided.

[0011] Andrea Simpson, Adam A. Hersbach, and Hugh J. McDermott, Improvements in speech perception with an experimental nonlinear frequency compression hearing device, International Journal of Audiology, 44:5:281-292, 2006 discloses further method of frequency mapping. The methods don't improve sound quality but enable the hearing-impaired to receive at least some information contained in the high frequency speech components.

[0012] Some further approaches are:

- to shift the complete spectrum to lower frequencies if high frequency sounds are detected (H. J. McDermott, V. P. Dorkos, M. R. Dean, and T. Y. Ching. Improvements in speech perception with use of the avr transonic frequency-transposing hearing aid. J Speech Hear Lang Res, 42(6):1323-1335, 1999). Problems with this scheme are the reliable detection of high frequency signals under noisy conditions and artifacts that occur during the on/off switching.
- to detect a range of high energy and to shift that region downwards (Kuk, Korhonen, Peeters, Jessen, and Andersen. Linear frequency transposition: Extending the audibility of high frequency information. The Hearing Review, 2006). Shifting means here, to overlap the identified frequency interval with a lower band and to add the two spectra.

[0013] This can lead to blurring of vowel sounds.

- to compress frequencies above a threshold frequency (described above). The compression is always switched on, regardless of the current speech sound.

[0014] US 5 771 299 aims to compress or expand the spectral envelope of an input signal. The goal is to move signal information from a region of the spectrum that is outside of the audible range of hearing aid users into another range that is still audible for the user. The system uses an LPC analysis filter and transmits the filter coefficients to the synthesis filter directly. Then, with the help of all-pass filters non-integer delays are introduced in the analysis and/or the synthesis filters. Delays larger than 1 compress the spectral envelope while delays smaller than 1 expand the spectral envelope.

[0015] The main problem with this system is that the voice specific characteristics such as formant positions are not preserved, since they are also compressed/expanded as part of the signal processing. This has the effect of enhancing audibility for the hearing impaired, but not of preserving the audio quality of the original input speech signal. Further disadvantages are delays of the output signal.

[0016] US 2009 0074197 discloses a similar transposition but with user-dependent information for spatial hearing. The goal is to perform a frequency transposition that moves frequency regions of the input signal to user specific ranges that are measured using built-in mechanisms of the hearing aid.

[0017] A method according to EP 1 333 700 A2 applies a perception-based nonlinear transposition function to the input spectrum. The same transposition function is applied over the whole signal so that artifacts resulting from switching between non-transposition and transposition processing are avoided. However, the use of a single function over the entire speech signal ignores time-varying phoneme-specific characteristics in the signal. Also, transforming the input signal to and then from the perception based scale to apply the transposition function reduces spectral resolution. As a result, characteristics of parts of the spectrum, which would otherwise only be subjected to a linear section of the transposition function, are not accurately preserved.

[0018] US 2006 0226016 aims to correct the phase of a transpositioned spectrum, wherein the transposition system is always active. Such processing of voiced segments has a negative effect on the phase and harmonic characteristics of the signal.

[0019] According to US 2009 0226016 the input signal is subjected to a high pass (or band pass) filter, after which the spectral envelope of the high pass signal is estimated using an all-pole model. A warping function is then applied to the all-pole model, translating the poles to lower frequencies using both non-linear and linear warping factors. An excitation signal (the prediction error signal) is then applied to the newly transposed all-pole model and shaped according to this transposed spectral envelope to create a synthesized transposed signal. The synthesized signal with an optional amplification and the original low-pass signal are then summed to create a signal containing compressed and transposed segments of the original spectrum. The LPC analysis and voice synthesis step is costly and the quality of such signals usually sounds quite artificial without proper postprocessing. This postprocessing is also expensive and this method is therefore not suitable for a system focusing on improved audio quality compared to systems for the hearing-impaired.

[0020] US 6 577 739 describes a proportional compression factor to the frequency spectrum generated by an FFT of the input signal. The factors are set to be between 0.5 and 0.99, and are applied using different methods. Using linear interpolation and assuming a compression factor of 0.5, the contribution of two input FFT bins are calculated and applied to one output IFFT pin. Another method would use an IFFT length double that of the input FFT length. By additionally placing the contributions from the input FFT bins in a higher or lower position in the output IFFT vector, a frequency transposition can be applied as well. This method again places a compressed frequency range of an input signal in another range in the output signal, which is still available to the communication channel or the end user of a hearing aid device. This system is however quite limited in scope by requiring FFT processing of the input and output signals. The additional time domain trimming of signals required by using FFT and IFFT vectors of different lengths also presents problems for time-domain synchronization and variability of the compression factors is not easily implemented.

[0021] CA 2 569 221 describes an attempt to retain information in frequency ranges outside of a band pass threshold. The input signal is converted to the frequency domain via the FFT algorithm or a polyphase filterbank. A compression function is then applied. Additionally, an amplification is applied in accordance to the energy level of the uncompressed portion of the input signal. The system can also comprise sending the compressed signal to an automatic speech recognition module. This method can have negative compression effects by giving portions of the input spectrum either too much or too little emphasis.

[0022] WO 2008/0123886 first defines a pass band for the output signal, and then defines a threshold, generally lower than the pass band, above which the frequency compression is applied. In the frequency region above the threshold, the highest frequency still of interest is identified and the appropriate compression is then applied. After the compression, the new peak power is normalized in a manner proportional to the compression, or is simply halved by -3 dB. Additionally, this method provides for the ability to expand a received signal, compressed or otherwise and synthetically reconstruct high frequency portions of the signal that may or may not have been present in the original transmitted signal. This can have a negative effect on the audio quality of the signal. This method still neglects the time-varying speaker dependent characteristics of speech sounds and applies the same compression function to each frame.

Summary of the Invention

[0023] In view of the foregoing, the need exists for improved solutions that preserve and enhance voice specific characteristics. The object of the present invention is to improve at least one out of controllability, precision, signal quality, processing load, and computational complexity.

[0024] The inventive method for adaptive spectral transformation for acoustic speech signals comprises the steps of receiving at least one spectral input representation corresponding to at least one window of a time domain input signal of acoustic speech,
selecting of the spectral input representations at least one selected spectral representation to be transformed,
assigning the at least one selected spectral representation to one of a set of cluster centres, wherein
the cluster centres are defined on the bases of spectral representations of windowed acoustic speech segments of a speech corpus by a clustering algorithm,
spectral class representations are assigned to the cluster centres and are elements of a code book and
the code book links to each spectral class representation at least one spectral transformation which enhances the corresponding spectral class representation,
transforming each selected spectral representation to a spectral output representation, wherein the applied transformation corresponds to the at least one spectral transformation linked to the cluster centre which is assigned to the respective selected spectral representation, and
providing the spectral output representations to synthesize an acoustic speech signal.

[0025] Selected spectral representations are classified by assigning them to spectral class representation. The spectral transformation applied to a selected spectral representation is adapted to this selected spectral representation because the applied spectral transformation enhances the assigned spectral class representation. The adaptations to enhance the spectral class representations are made in a setup procedure and include heuristic steps in order to find transformations which enhance the spectral class representations. The adapted advantageous transformations for each cluster centre respectively for each class of the code book ensure enhancement of the spectral representation of all the spectral representations assigned to this class of the code book. The controllability, precision and signal quality are enhanced while the processing load and computational complexity are reduced.

[0026] Best results can easily be found by the use of an appropriate code book, respectively by selecting an appropriate speech corpus and an appropriate clustering algorithm. Clustering of spectral representation with sufficient energy in the 4 to 8 kHz band allows specific enhancement of fricatives such as /s/, /sh/, /ch/, /z/ or /f/. The step of finding transformations which enhance fricatives of different classes allows linking of very specific transformations to the classes of the code book.

[0027] An English speech corpus can for example be taken from the TIMIT acoustic-phonetic continuous speech data base. This speech data base has been designed by the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI) for acoustic-phonetic studies. It contains utterances of 630 male and female speakers in American English, divided into eight dialect regions. From each person, ten phonetically rich sentences have been recorded and tagged with phonetic information. The training data can be reduced to 70 speakers from all dialect regions and pauses can been removed with the help of phoneme tags. The reduced training data has a duration of approximately 23 minutes. For validation purposes, a second data set can been extracted that consists for example of 10 minutes speech data from 30 speakers. The sampling frequency is preferably fs = 16 kHz.

[0028] There are two different kinds of preferred transformations, frequency compression and formant boosting. Frequency compression is compressing the bandwidth for example from a bandwidth of 0 to 8 kHz to an bandwidth of 0 to 4 kHz preferably with a compression only at the upper or lower end of the bandwidth, optionally linear at least in the middle frequency range, corresponding to no compression when the slope is equal to 1. Formant boosting takes place using formant-dependent functions to increase the contrast between the formants and the non-formant frequencies in the frequency spectrum. The formant boosting gain function linked to each spectral class representation of the code book is amplifying at least one preferably two or three of the formants at low frequencies.

[0029] Assigning the at least one selected spectral representation to one of a set of cluster centres includes calculating distance measures between the selected spectral representation and all the
spectral class representations of the code book, and
assigning the at least one selected spectral representation to the cluster centre with the shortest distance measures between the selected spectral representation and the spectral class representations of the cluster centre.

[0030] Calculating distance measures becomes very simple by first calculating feature vectors to the spectral representations. The distance measures are just distances between the feature vectors. The feature vectors are calculated from spectral envelope representations by a filtering transformation, preferably with a mel-filterbank, wherein the mel-filterbank optionally uses overlapping triangular windows with widths variable with frequency. With such feature vectors it can be sufficient to use a code book including at least eight (for example for fricative enhancement), preferably thirty-two, optionally 128 cluster centres. With higher class numbers more different enhancement problems can be solved each in a different way by the linked at least one spectral transformations.

[0031] The spectral class representations for the cluster centres are preferably averaged spectral representations averaged over spectral representations of corresponding cluster elements. A reduction to spectral class representations of special interest can be made by applying the clustering algorithm on the bases of a preselected sub-corpus of the speech corpus. The sub-corpus can for example be reduced to spectral representations of the speech corpus which have a spectral centroid lying above a given threshold frequency, preferably of 3 kHz.

[0032] Transformations are only needed for spectral representation which can be improved. A selection can be made by calculating the spectral centroid of each spectral input representation and selecting spectral input representations which have a spectral centroid lying above a threshold, frequency preferably above 3 kHz. Such a selection fits to a code book base on a sub-corpus of the speech corpus with the same threshold frequency. A selection can also be made by detecting at least one of speech activity and background noise level. Spectral input representations with speech to be transformed will be selected with appropriate selection criteria.

[0033] The method for adaptive spectral transformation for acoustic speech signals can be implemented in different fields. In some applications the acoustic speech signal is windowed and of each window a spectral representation is deduced. There are also applications where the spectral representations of windows of an acoustic speech signal are provided by a system. Therefore the method of this invention starts with the step of receiving spectral input representations, which can be provided by the same system or by another system. At least some of the received spectral input representations are selected and enhanced by adapted transformations and the enhanced spectral representations are provided in the form of at least one spectral output representation. The combination of untransformed and transformed spectral representations allow synthesizing an enhanced acoustic speech signal.

[0034] The invention can be implemented in a computer program comprising program code means for performing all the steps of the disclosed methods when said program is run on a computer.

[0035] The described solutions preserve and enhance voice specific characteristics such as formants and their respective positions. This has the effect of enhancing audibility, while preserving the audio quality of the original input speech signal. The current invention also avoids unnecessary delays of the output signal. Various compression functions are applied to the speech signal, which takes into account the time-varying phoneme-specific characteristics in the signal and allows for spectral sharpening through adaptive formant boosting. Also, unnecessary transformations of the input signal, which could reduce spectral resolution are avoided. The effect of phase distortion on the harmonic section of the input signal can be avoided in the current invention by limiting the compression processing to only unvoiced segments of speech.

[0036] A focus on efficient algorithms and output signal synthesis allows the current invention to also avoid costly postprocessing to achieve high quality audio output. Additionally, transforming the input signal from the time domain to the frequency domain is not bound to the FFT (Fast Fourier Transform) algorithm. Compression functions are intended to be designed such that they are able to be efficiently stored in memory and do not inadvertently give portions of the input spectrum undesired emphasis.

Brief description of the figures

[0037]

Fig. 1 spectrogram of the utterance "Sauerkraut is served once a week". The plots below show the percentage of signal energy above 4 kHz and the amplitude of the signal

Fig. 2 example of the LBG-algorithm for clusters of two dimensional feature vectors with (c) k = *K* = 8 classes, x- and y-axis correspond to the value ranges of the first and the second features

Fig. 3 results of clustering with K = 8 in classes after pre-classification with a threshold of *fT* = 3kHz. The line with more details shows the average spectrum in dB, the line with less details shows a filtered spectral representation of the code book entry and the vertical line the spectral centroid fc

Fig. 4 a block diagram of the adaptive transformation method

Fig. 5 and 6, compression functions

Fig. 7 a formant boosting gain function and spectral representations (cluster mean and filtered cluster mean), where the line with more details shows the average spectrum in dB, the line with less details shows a filtered spectral representation of the code book entry

Fig. 8 a block diagram of the adaptive transformation method using feature vectors

Fig. 9 examples for processing of phonemes: (a) spectral compression with compression characteristic below, (b) formant boosting with gain $g(\Omega_\mu)$ between + -10dB. The processed signals are band limited to 4 kHz (shorter curve)

Fig. 10 example of the sauerkraut utterance processed with the 128 class scheme (input signal above, processed signal below)

Fig. 11 time series auf a noise reduction filter and its derivative

Fig. 12 block diagram of the onset sharpening using an algorithm according to $G_{rec}^{(mod\ 1)}(\Omega_\mu, l)$ of the recursive Wiener filter

Detailed description of the invention

[0038] Fig. 1 shows for the acoustic speech "Sauerkraut is served once a week" the energy distribution in time and frequency, a time series of the energy above 4 kHz and a time series of the amplitude. The fricatives "s", "ce" and "k" have quite some energy in the frequency range from 4 to 8 kHz.

[0039] A time-domain input signal, as for example "Sauerkraut is served once a week", is windowed before being transferred to the frequency domain via a fast Fourier transform (FFT) algorithm, discrete Fourier transform (DFT), discrete cosine transform (DCT), polyphase filterbank, wavelet transform, or some other time to frequency domain transformation. In the frequency domain the windowed time-domain signal has the form of spectral representations. By filtering for example with a mel-filterbank the spectral representations can be reduced to feature vectors.

[0040] Fig. 2 shows an example of clusters of two dimensional feature vectors. With the LBG-algorithm for (c) k = *K* = 8 classes eight cluster centers + are found. The procedure is similar for feature vectors with higher dimensionality. The spectral representations or the feature vectors of the cluster centres are defined on the bases of spectral representations of windowed acoustic speech segments of a speech corpus by the clustering algorithm. A code book includes the spectral class representations for the cluster centres, which are averages over the elements of the cluster.

[0041] Fig. 3 shows results of clustering with K = 8 in classes after pre-classification with a threshold of *fT* = 3kHz in (b). The line with more details shows the average spectrum in dB, the line with less details shows a filtered spectral representation of the code book entry and the vertical line the spectral centroid *fc* . The filtered spectral representations of the code book entries are the spectral class representations for the cluster centres, which are averages over the elements of the cluster.

[0042] The spectral representations or feature vectors of windowed frames of an input signal are subjected to a classification technique from the field of pattern recognition, such as the minimum mean squared error estimator. The input frames are classified by finding the class corresponding to the smallest value of a cost function, $d(\mathbf{v},\mathbf{c}_k)$ in Equation 1, where v_d is the D-dimensional set of features of the current frame and c_{dk} is the set of features of a codebook entry k.

Equation 1

$$d(\mathbf{v}, \mathbf{c}_k) = \sum_{d=1}^{D} (v_d - c_{dk})^2$$

[0043] In Fig. 4, this setup is illustrated using K=128 as a possible number of entries for such a codebook. The codebook can be trained using feature vectors from a training set and any number of vector quantization algorithms, such as k-means [MacQueen, 1967] or the Linde-Buzo-Gray (LBG) algorithm[Linde, Buzo, & Gray, 1980]. Linear discriminant analysis (LDA), principal component analysis (PCA), support vector machines (SVM), and other class enhancing algorithms can also be applied to decrease the intraclass variances and/or increase the interclass distances, which improves separability.

[0044] The feature vectors for training and testing are in this case a set of perception based features in the mel scale, although the feature vectors must not be limited to these specific features. The features can also be designed to emphasize signal characteristics in or near the region of interest in the frequency spectrum. The key concept here is that the feature vector has a reduced dimensionality with respect to the input signal frequency spectrum in order to reduce computational costs.

[0045] A spectral compression function and/or a formant boosting function are chosen from the relevant codebook entry.

[0046] To avoid possible switching artifacts, a smoothing procedure can be applied to the chosen compression function in the frequency domain but in the time direction. Using for example, an IIR-smoothing function, the time-variance of the applied compression functions can be reduced.

[0047] The spectral compression functions are functions- in the preferred case continuous and nonlinear - that apply compression rates in the frequency domain. Both the compression functions and the frequency spectrum can be either linearly or nonlinearly scaled, reflecting the occasional advantages of processing in more perception based scales like the logarithmic scale.

[0048] The operation of spectral compression maps a frequency interval of the input signal into a smaller frequency interval of the output. When working on a discrete representation of the spectrum, this is equivalent to mapping a set of frequency pins from the input spectrum X($e^{j\Omega\mu}$) into one frequency pin of the output spectrum Y ($e^{j\Omega\mu}$). Mathematically, an operation that reduces the input frequency bandwidth by approximately 0.5 can be expressed by

Equation 2

$$Y(e^{j\Omega_\nu}) = \frac{X(e^{j\Omega_\nu})}{\left|X(e^{j\Omega_\nu})\right|} \frac{1}{\mu_u^{(\nu)} - \mu_l^{(\nu)} - 1} \sum_{\mu=\mu_l^{(\nu)}}^{\mu_u^{(\nu)}} \left|X(e^{j\Omega_\mu})\right| \qquad \nu = 0, \ldots, \frac{N}{4}+1.$$

[0049] The upper and lower boundaries of the interval that is projected onto output frequency pin *v* are represented by the variables $\mu_u^{(v)}$and $\mu_l^{(v)}$, respectively. These boundaries are actually functions of *v*, allowing a variable amount of compression for different frequencies. Compression with this equation derives the magnitude of Y ($e^{j\Omega v}$) as the mean value of the magnitude of X($e^{j\Omega\mu}$) for $\mu = \mu_u^{(v)}, \ldots , \mu_l^{(v)}$ while retaining the phase of input component *v* for output component *v*.

[0050] Fig. 5 illustrates an example for the relationship between input frequency f_{in} and output frequency f_{out}. The curved line shows the compression characteristic, the horizontal and vertical lines indicate the frequency interval that is mapped. The frequency in Hz can be converted to the pin-index with the following relationship

Equation 3

$$\mu = \left\lfloor \frac{f}{f_s} N_{\mathrm{FFT}} \right\rceil$$

[0051] Equation 2 also performs energy normalization with respect to the amount of frequency compression. Energy normalization can take many forms, however, and could also be applied in a fashion based on the momentary broadband or frequency localized SNR. Another option is to maintain the form of the precompression estimated spectral envelope and apply a gain or dampening factor to correct the energy level to correspond to the slope of the envelope in the region of interest.

[0052] The equation also preserves the phase of the original uncompressed signal, but other phase corrections and adjustments are also possible, such as using a random phase.

[0053] The compression functions can be continuous in nature and have various compression rates over frequency. Preferably, the compression function is linear in lower frequencies, corresponding to no compression when the slope is equal to 1. In the higher frequencies of interest, the compression rate increases gradually, with the rate and degree of increase dependent upon the feature-based classification.

[0054] In Fig. 6 various examples of continuous nonlinear compression functions are shown. Those displayed with solid lines are linear with a slope of, or near, 1 in the lower frequency range, hence performing little to no compression. With increasing input frequency, the compression rate increases differently for each of the curves. The dashed curves perform little to no compression in the middle frequency range, rather than the low frequencies. These curves compress frequencies in the lower and higher frequency extremes, and can be better understood as performing a transposition of the middle frequencies down into a lower region with almost no compression.

[0055] The appropriate compression function for each class of the code book is to be defined either manually, e.g., using subjective listening criteria, or automatically, e.g., applying unsupervised learning methods to find a preferred mapping to an output characteristic. Different sets of class function assignments are generally possible, since the ideal output for improved audio quality does not always coincide with that providing improved speech recognition rates.

[0056] Speech can be made more accentuated, if formants are amplified. Especially in situations with background noise, we can expect a better localized SNR in these frequency regions, so the broadband SNR can be improved by applying a frequency dependent gain factor to the input spectrum. Ideally, the amplification should only be applied during speech activity. Otherwise, background noise will be amplified during pauses.

[0057] The current invention can receive information about speech activity from an external module, such as a noise detection and cancelation module.

[0058] The formant-boosting functions are also functions - (non)continuous, (non)linear, and on a (non)linear scale - designed such that a variable gain factor is applied to frequency ranges around formants in the spectrum. Fig. 7 shows a formant boosting gain function and spectral representations (cluster mean and filtered cluster mean), where the line with more details shows the average spectrum in dB, the line with less details shows a filtered spectral representation of the code book entry. A curve can be used to determine the percentage of the gain factor that is applied to the formant frequency range. Alternatively individual gain and dampening curves can be stored in a codebook and applied to those frequency ranges identified as formants in voiced speech.

[0059] For some spaces between formants, the gain factor just described can be transformed into a dampening factor using another set of curves. The boosting curves can be designed to have positive values for amplifying the formant frequencies and negative values for the dampening curves to reduce the magnitude of the valleys between formants.

[0060] When compression is applied an additional curve can be used in the frequency range of the compression results. This amplifies or dampens the spectrum in the region of the frequency compression and so can be used to amplify or dampen the effects of the compression.

[0061] A decision can be made for each codebook entry of the spectral compression, formant boosting signal processing about the order of the steps. In some instances it can be necessary to first apply spectral compression and then the boosting function. At other times it should be possible to first boost (or dampen) regions of the spectrum and then to apply the spectral compression. This could be important when the compression is designed such that one of the formants is located in the compressed frequency range. The prior boosting of the formant would serve to retain the formant shape even after compression.

[0062] An overall block diagram of one possible incarnation of the system is seen in Fig. 8. Windows of a signal in the time domain are transformed to spectral representations by an analysis filter-bank. By extracting feature vectors a classification in relation to elements of a code book a signal processing is realized with the transformations linked to the code book elements. The transformed spectral representations are transformed to windows of a signal in the time domain by a synthesis filter-bank.

[0063] Fig. 9 shows examples for processing of phonemes:

(a) with a spectral compression according to the shown compression characteristic ,
(b) with formant boosting according to the shown gain $g(\Omega_\mu)$ between + -10dB. The processed signals are band limited to 4 kHz.

[0064] Fig. 10 shows an example of the sauerkraut utterance processed with the 128 class scheme. The energy of the fricatives "s", "ce" and "k" is transformed below 4 kHz.

[0065] The system is also capable of receiving information from other signal processing modules, such as silence/ unvoiced/voiced decisions from the noise estimation and reduction module. Furthermore the module should be capable of sending information to other modules, especially an ASR module, which can use the enhanced signal to improve recognition rates. The ASR module could be retrained with the compressed output data. However the compressed signal alone achieves a change in recognition rates that can be useful.

[0066] A focus on efficient algorithms and output signal synthesis allows the current invention to also avoid costly postprocessing to achieve high quality audio output. Additionally, transforming the input signal from the time domain to the frequency domain is not bound to the FFT algorithm. Compression functions in the current invention are intended to be designed such that they are able to be efficiently stored in memory and do not inadvertently give portions of the input spectrum undesired emphasis.

Onset Sharpening

[0067] Another invention is disclosed, which is new and inventive independent of the independent claims. This further invention is related to onset sharpening, which can be used independently but is of course advantageously combinable with the previously described speech enhancement methods.

[0068] The onset sharpening method performs an onset sharpening, i.e., to introduce attenuation immediately before and/or amplification after speech onsets in order to make them more accentuated. This is improving speech quality and intelligibility, especially for speech signals corrupted with background noise that are to be enhanced with noise reduction methods. Noise reduction filters and their derivatives (Fig. 11) tend to react too slow and therefore do remove desired signal components during speech onsets.

[0069] Before any onset-dependent signal enhancement can be performed, the speech onsets need to be found first. This is done based on a recursive Wiener noise reduction filter. There are no real-time constraints, so all of the following steps can be applied to the entire signal x(t) giving the necessary data for the next step.

[0070] An analysis filter bank is needed to transform the input signal x(t) into the frequency domain. The result is a function $X(e^{j\Omega\mu}, l)$ with $\mu = 0, \ldots, N_{DFT}=2$ and $l = 0, \ldots, M-1$, where M is the number of signal frames. This could also be interpreted as a $(N_{DFT}=2+1) \times M$ matrix which is constituting a spectrogram.

[0071] Based on the spectral signal representation of the previous step, the attenuation factors of the recursive Wiener filter characteristic can be computed. The following, non-frequency dependent, parameters have been used as filter parameters:

- Maximum attenuation $\underset{\sim}{G}_{min}({}^{\Omega}\mu, \underset{\sim}{l}) = G_{min} = 0.25$ (corresponding to $20 \log_{10}(G_{min}) = -20$ dB)
- Overestimation factor $\beta(\Omega_\mu, l) = \beta = 5$

[0072] The result $G_{rec}(\Omega\mu, l)$ can again be seen as a $(N_{DFT}=2+1) \times M$ matrix containing the attenuation factor for each sub-band for all time instances.

[0073] In order to smooth the filter coefficients in a perceptually meaningful manner over time, a mel-filterbank of 32 bands is applied to $G_{rec}(\Omega_\mu, l)$, resulting in the 32 x M matrix $G_{rec}^{(mel)}(m, l)$. The actual onset detection is performed within each mel-band of $G_{rec}^{(mel)}(m, l)$.. Here, the moments when $G_{rec}^{(mel)}(m, l)$ changes its value from G_{min} to 1 (or close to 1) are of interest, i.e., the points when the filter opens. These time instances can be found by taking the numerical derivative

$$\frac{dG_{rec}^{(mel)}(m,l)}{dl} \approx G_{rec}^{(mel)}(m,l) - G_{rec}^{(mel)}(m,l-1)$$

and comparing the resulting value with a threshold

$$\frac{dG_{rec}^{(mel)}(m,l)}{dl} > \gamma$$

[0074] The mel-band m for time instance l is labeled as a speech onset. The derivative of the noise reduction coefficient lies in the range of d $G_{rec}^{(mel)}$(m,l)/dl ∈ [G_{min} -1, 1- G_{min}] and positive values indicate times when the filter is opening. A threshold = 0.2 has proved to give good detection results for various speech signals and SNRs. Because it could happen that the derivative is greater than for several consecutive frames, also a sliding time window of 100ms duration is applied. Within this time window, only one detection is allowed. Furthermore, if an onset has been detected in a certain mel-band, the neighboring mel-band towards lower frequencies of the same frame l is also marked as a speech onset. The clean speech is mixed with background noise recorded in a car driving at a speed of 160 km/h to form an SNR of 1 dB during speech activity. Of course, the detection can be made more sensitive by taking a lower threshold, e.g. y= 0.1.

[0075] Based on the detections made with the method of the previous section, the onset sharpening can be performed. It consists of placing attenuation immediately before a detected speech onset and boosting the signal for a short time interval afterwards. The shape of the attenuation/amplification is determined by a prototype function that has been chosen to be

$$f(x) = \begin{cases} -\varrho \frac{e}{\sigma} x e^{\frac{x}{\sigma}} \, for \, -20 \le x < 0 \\ \frac{e}{\sigma} x e^{\frac{-x}{\sigma}} \, for \; 0 \le x \le 20 \\ 0 \; otherwise \end{cases}$$

with the attenuation ϱ = 0.5 and σ = 3. The term e/ σ is used to normalize f (x) to a maximum value of 1. Out of this prototype function, the onset sharpening gain function g_{os}(l) can be sampled. When deriving onset sharpening gain function, three parameters can be set:

1. The width of the (negative) attenuation and the (positive) boosting part, defined by the variable τ_{os} in ms.
2. The offset τ_{offs} in ms that defines at which time instance the gain function is placed. For τ_{offs} = 0 ms, the zero crossing is exactly at the detection point, for negative offset values the zero crossing will be earlier. This is desirable because then the noise reduction filter is forced to open earlier.
3. A gain parameter α_{os} that controls the amount of attenuation and boosting. It is multiplied with the onset sharpening gain function g_{os}(l) which is interpreted as dB values.

[0076] Several other prototype functions could be devised for onset sharpening, e.g., a sinusoid. The prototype function has been chosen with the stated parameters because it decays smoothly towards the ends and offers a steep slope around the zero crossing. This prototype function is now placed at each detected speech onset time instance in the corresponding mel-band, giving the onset sharpening matrix for mel-bands $g_{os}^{(mel)}(m,l)$, which then is expanded into the onset sharpening matrix for the subbands g_{os}(m,l). The parameters that have been used for the onset sharpening gain functions are τ_{os} = 75 ms, τ_{offs} = -10ms and α_{os} = 3, of course other parameters are possible as well. Using the notation of the mel-filterbank matrix A the expansion from melbands to subbands can be expressed as

$$g_{os}(\Omega_\mu, l) = \sum_{m=0}^{D} \frac{A_{m\mu}}{max_\mu\{A_{m\mu}\}} * g_{os}^{(mel)}(m, l).$$

[0077] The weighting of the filters contained in A that gives triangles of a broader bandwidth a lower amplitude is removed by the normalization containing the maximum operation.

[0078] Fig. 12 discloses an onset sharpening algorithm with the following recursive Wiener noise reduction filter being modified by the onset sharpening gain function. The simplest way is by multiplication

$$\begin{aligned} G_{rec}^{(mod\ 1)}(\Omega_\mu, l) &= g_{os}^{lin}(\Omega_\mu, l) \cdot G_{rec}(\Omega_\mu, l) \\ &= g_{os}^{lin}(\Omega_\mu, l) \cdot \max\left\{G_{min}(\Omega_\mu, l), 1 - \frac{\tilde{\beta}(\Omega_\mu, l)}{G(\Omega_\mu, l-1)} \cdot \frac{\left|\hat{B}(e^{j\Omega_\mu}, l)\right|^2}{\left|\hat{X}(e^{j\Omega_\mu}, l)\right|^2}\right\} \end{aligned}$$

where

$$g_{os}^{lin}(\Omega_\mu, l) = 10^{\alpha_{dB} \cdot g_{os}(\Omega_\mu, l)/20}$$

is the onset sharpening function in linear values. Since the noise reduction filter is applied multiplicative to the input spectrum, this filter modification could also be interpreted as a multiplication of the signal spectrum $X(e^{j\Omega_\mu}, l)$ with the onset sharpening gain before (or after) applying the noise reduction filter.

[0079] In a second modification, the onset sharpening gain is built into the noise reduction characteristic to modify the spectral floor and the maximum gain of the filter (which is set to 1 in the recursive Wiener filter):

$$G_{rec}^{(mod\ 2)}(\Omega_\mu, l) = \max\left\{G_{min}(\Omega_\mu, l) \cdot g_{os}^{att}(\Omega_\mu, l), g_{os}^{amp}(\Omega_\mu, l) - \frac{\left|\hat{B}(e^{j\Omega_\mu}, l)\right|^2}{\left|\hat{X}(e^{j\Omega_\mu}, l)\right|^2}\right\}.$$

[0080] For this description, the gain function has to be separated into the part responsible for the attenuation before speech onsets

$$g_{os}^{att}(\Omega_\mu, l) = \begin{cases} 10^{\alpha_{dB} \cdot g_{os}(\Omega_\mu, l)/20} & for\ g_{os}(\Omega_\mu, l) < 0 \\ 1 & \text{otherwise} \end{cases}$$

and the part for amplification after speech onsets

$$g_{os}^{amp}(\Omega_\mu, l) = \begin{cases} 10^{\alpha_{dB} \cdot g_{os}(\Omega_\mu, l)/20} & for\ g_{os}(\Omega_\mu, l) \geq 0 \\ 1 & \text{otherwise} \end{cases}$$

[0081] A third possibility is to modify also the overestimation factor:

$$G_{\mathrm{rec}}^{(\mathrm{mod}\ 3)}(\Omega_\mu,l) = \max\left\{G_{\min}(\Omega_\mu,l)\cdot g_{\mathrm{os}}^{\mathrm{att}}(\Omega_\mu,l),\, 1-\frac{\tilde{\beta}(\Omega_\mu,l)}{g_{\mathrm{os}}^{\mathrm{amp}}(\Omega_\mu,l)\cdot G(\Omega_\mu,l-1)}\frac{\left|\hat{B}\,(e^{j\Omega_\mu},l)\right|^2}{\left|\hat{X}\,(e^{j\Omega_\mu},l)\right|^2}\right\}$$

[0082] Inspection of noise reduction filters from the first and second modification shows that there are only small differences between the characteristics. The parameter settings for τ_{os}, τ_{offs}, α_{os} and the choice of the gain prototype function are of much greater influence. Therefore, the simpler modification $G_{rec}^{(mod\ 1)}(\Omega_\mu,l)$) will be used for the evaluation.

[0083] For the evaluation of the speech onset enhancement method, the recursive Wiener filter with $G_{rec}^{(mod\ 1)}(\Omega_\mu,l)$ has been compared to a standard recursive Wiener filter $G_{rec}(\Omega_\mu,l)$. This has mainly been done on the basis of a logarithmic spectral distance (LSD) measure. Comparison of the noise reduction filter coefficients for several characteristics gives a qualitative impression about the opening/closing properties of a filter. Finally, listening tests give a useful criterion that help to judge intelligibility and the amount of artifacts such as musical tones.

[0084] The idea in using an LSD measure is to create a signal x(t) = s(t)+ b(t) corrupted with background noise, where the speech component s(t) and the noise b(t) are known. A noise reduction filter is computed for the disturbed signal x (t) and the two signal components are passed through this filter separately. Then, the distortion measures LSD_{speech} and LSD_{noise} can be calculated between the original and the filtered signal. Ideally, the speech component is passed through the filter unchanged, leading to a distance close to zero, whereas large distortions can be expected for the noise component. Based on these two measures, it is possible to judge on the noise suppression ability and on how much speech components are affected. The LSD between two time varying spectra $S(e^{j\Omega_\mu}, l)$ and $\hat{S}(e^{j\Omega_\mu}, l)$ is defined as

$$LSD = \frac{10}{D}\sum_{l=0}^{L}\sqrt{\sum_{\mu=0}^{N_{DFT}/2}\frac{K_{\mu,l}}{\overline{K}_l}\log_{10}^2\left\{\frac{\max\left\{\left|S(e^{j\Omega_\mu},l)\right|^2,\delta_S\right\}}{\max\left\{\left|\hat{S}(e^{j\Omega_\mu},l)\right|^2,\delta_{\hat{S}}\right\}}\right\}}$$

[0085] The two variables

$$\delta_S = 10^{-5}\max_{\mu,l}\left\{\left|S(e^{j\Omega_\mu},l)\right|^2\right\}$$

$$\delta_{\hat{S}} = 10^{-5}\max_{\mu,l}\left\{\left|\hat{S}(e^{j\Omega_\mu},l)\right|^2\right\}$$

give lower bounds for the values that enter the measure. These components are selected by the binary mask

$$K_{\mu,l} = \begin{cases} 1\ if\ \left|S(e^{j\Omega_\mu},l)\right|^2 \geq \delta_S \\ 0\ otherwise \end{cases}$$

[0086] The normalization factor $\overline{K}_l$ counts the number of components that are used for the distance measure in each time frame. In order to avoid a division by zero, it is defined as

$$\overline{K}_l = \max\left\{ \sum_{\mu=0}^{N_{DFT}/2} K_{\mu,l}, 0.1 \right\}$$

[0087] The variable D gives the number of signal frames that are used for the calculation of the LSD, i.e., the number of signal frames with $\overline{K}_l > 0$.

[0088] For evaluating the performance of the modified recursive Wiener filter $G_{rec}^{(mod\ 1)}(\Omega_\mu, l)$,, a set of 616 filters has been computed with all possible combinations of the parameters $\tau_{os} \in [50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100]$ ms
$\tau_{offs} \in [0, -5, -10, -15, -20, -25, -30, -35]$ ms
$\alpha_{os} \in [0, 2, 4, 6, 8, 10, 12]$.

[0089] The signal that has been used is the "Sauerkraut is serve once a week" utterance used throughout this text, mixed with background noise from a car driving at a constant speed of 160km/h. The SNR d ring speech activity is adjusted to 1 dB. Then, only the speech and only the noise components have been processed with these filters and the distances $LSD_{speech}^{(os)}$ and $LSD_{noise}^{(os)}$ have been calculated. As reference for the comparison, a recursive Wiener filter has been designed and the measures $LSD_{speech}^{(rw)}$ and $LSD_{noise}^{(rw)}$ have been evaluated.

[0090] As mentioned earlier, a good filter is characterized by a small LSD for speech and a large value for noise. Obviously, these two requirements are difficult to meet at the same time and an increase in one category usually falls together with an increase in the other one. For better comparison of the two filter types, the $LSD_{speech}^{(rw)}$- $LSD_{speech}^{(os)}$ and $LSD_{noise}^{(os)} - LSD_{noise}^{(rw)}$ could be calculated They are defined such, that a positive value means that the onset sharpening approach gives better results in the LSD sense. This is apparently not always the case and again it can be seen that an improvement in noise suppression leads to worse results for speech and vice versa. However, there are some combinations where a positive value can be achieved in both disciplines, e.g., for the parameter combination
$\tau_{os} = 100$ms
$\tau_{offs} = -30$ ms.

[0091] The gain factor α_{os} basically only scales the distance measure. For the special case of $T_{os} = 0$, the modified filter reduces to the standard recursive Wiener filter and thus gives the same LSD.

[0092] Comparing the noise reduction filter coefficients for a Wiener filter, a recursive Wiener filter and a recursive Wiener filter modified by onset detection for several subbands and different filter parameters gives the following result. The same signal as for the LSD measures has been used for the design of these filters.

[0093] The parameters that have been used are

$$[\tau_{os}, \tau_{offs}, \alpha_{os}] = [75\text{ms}, -10\text{ms}, 3]$$

[0094] But of course other parameters are possible, too.

[0095] The Wiener filter opens more often, which potentially results in musical tones. It has also been seen that the recursive Wiener filter opens a bit later, which can be corrected by the onset sharpening modification.

[0096] First of all it should be noticed that the offset τ_{offs} = -30ms is fairly large compared to the duration of τ_{os} = 100 ms. This causes that the modified filter coefficient near a speech onset increases, then decreases for a few frames and finally grows again with the opening of the recursive Wiener filter. Apparently, this is no optimal behavior even though this parameter combination gave the best results in the LSD. This gives rise to the assumption, that a different prototype function should be used. Good candidates would be more flat around their maximum, avoiding the "crash down" near 0.18 and 0.5 seconds. At any rate, a filter that is opening earlier seems to give benefits in terms of the LSD measure.

[0097] However, for larger gains α_{os}, the noise floor was decreased on the expense of increasing filtering artifacts.

[0098] Fig. 11 shows a procedure where the attenuation factor and its derivative together with the threshold are shown for mel-band 25 (corresponding to frequencies between 3.6 and 4.4 kHz). Because it could happen that the derivative is greater than y for several consecutive frames, also a sliding time window of 100ms duration is applied. Within this time window, only one detection is allowed. Furthermore, if an onset has been detected in a certain mel-band, the neighboring mel-band towards lower frequencies of the same frame l is also marked as a speech onset.

[0099] The signal that has been used is the utterance "Sauerkraut is served once a week" from the TIMIT database that has also been used before. The clean speech is mixed with background noise recorded in a car driving at a speed of 160 km/h to form an SNR of 1 dB during speech activity.

Claims

1. A method for adaptive spectral transformation for acoustic speech signals comprising the steps of
receiving at least one spectral input representation corresponding to at least one window of a time domain input signal of acoustic speech,
selecting of the spectral input representations at least one selected spectral representation to be transformed,
assigning the at least one selected spectral representation to one of a set of cluster centres, wherein
the cluster centres are defined on the bases of spectral representations of windowed acoustic speech segments of a speech corpus by a clustering algorithm,
spectral class representations are assigned to the cluster centres and are elements of a code book and
the code book links to each spectral class representation at least one spectral transformation which enhances the corresponding spectral class representation,
transforming each selected spectral representation to a spectral output representation, wherein the applied transformation corresponds to the at least one spectral transformation linked to the cluster centre which is assigned to the respective selected spectral representation, and
providing the one spectral output representations to synthesize an acoustic speech signal.

2. Method as claimed in claim 1, wherein assigning the at least one selected spectral representation to one of a set of cluster centres includes
calculating distance measures between the selected spectral representation and all the spectral class representations of the code book, and
assigning the at least one selected spectral representation to the cluster centre with the shortest distance measures between the selected spectral representation and the spectral class representations of the cluster centre.

3. Method as claimed in claim 2, wherein calculating distance measures includes calculating feature vectors for the spectral representations and the distances measures are distances between the feature vectors.

4. Method as claimed in claim 3, wherein the feature vectors are calculated from the spectral representations by a filtering transformation, preferably with a mel-filterbank, wherein the mel-filterbank optionally uses overlapping triangular windows with widths variable with frequency.

5. Method as claimed in one of claims 1 to 4, wherein the code book includes at least eight, preferably thirty-two, optionally 128 cluster centres.

6. Method as claimed in one of claims 1 to 5, wherein the spectral class representations for the cluster centres are averaged spectral representations averaged over spectral representations of corresponding cluster elements respectively classes.

7. Method as claimed in one of claims 1 to 6, wherein the definition of cluster centres by the clustering algorithm is made on the bases of a preselected sub-corpus of the speech corpus.

8. Method as claimed in claim 7, wherein preselecting a sub-corpus includes the reduction to spectral representations of the speech corpus which have a spectral centroid lying above a threshold frequency, preferably above 3 kHz.

9. Method as claimed in one of claims 1 to 8, wherein one of the at least one spectral transformation linked to each spectral class representation of the code book is a spectral compression transformation mapping a frequency interval of the selected spectral representation to a smaller frequency interval of the spectral output representation.

10. Method as claimed in claim 9, wherein the spectral compression transformation linked to each spectral class representation of the code book is compressing an bandwidth of 0 to 8 kHz to an bandwidth of 0 to 4 kHz preferably with a compression only at the upper or lower end of the bandwidth, optionally linear at least in the middle frequency range, corresponding to no compression when the slope is equal to 1.

11. Method as claimed in one of claims 1 to 7, wherein one of the at least one spectral transformation linked to each spectral class representation of the code book is a formant boosting gain function.

12. Method as claimed in claim 11, wherein the formant boosting gain function linked to each spectral class representation of the code book is amplifying at least one preferably two or three of the formants at low frequencies.

13. Method as claimed in claim 8, wherein selecting at least one selected spectral representation to be transformed includes calculating the spectral centroid of each spectral input representation and selecting spectral input representations which have a spectral centroid lying above a threshold, frequency preferably above 3 kHz.

14. Method as claimed in one of claims 1 to 13, wherein selecting at least one selected spectral representation to be transformed includes detecting at least one of speech activity and background noise level and selecting spectral input representations with speech to be transformed.

15. A computer program comprising program code means for performing all the steps of any one of the claims 1 to 14 when said program is run on a computer.

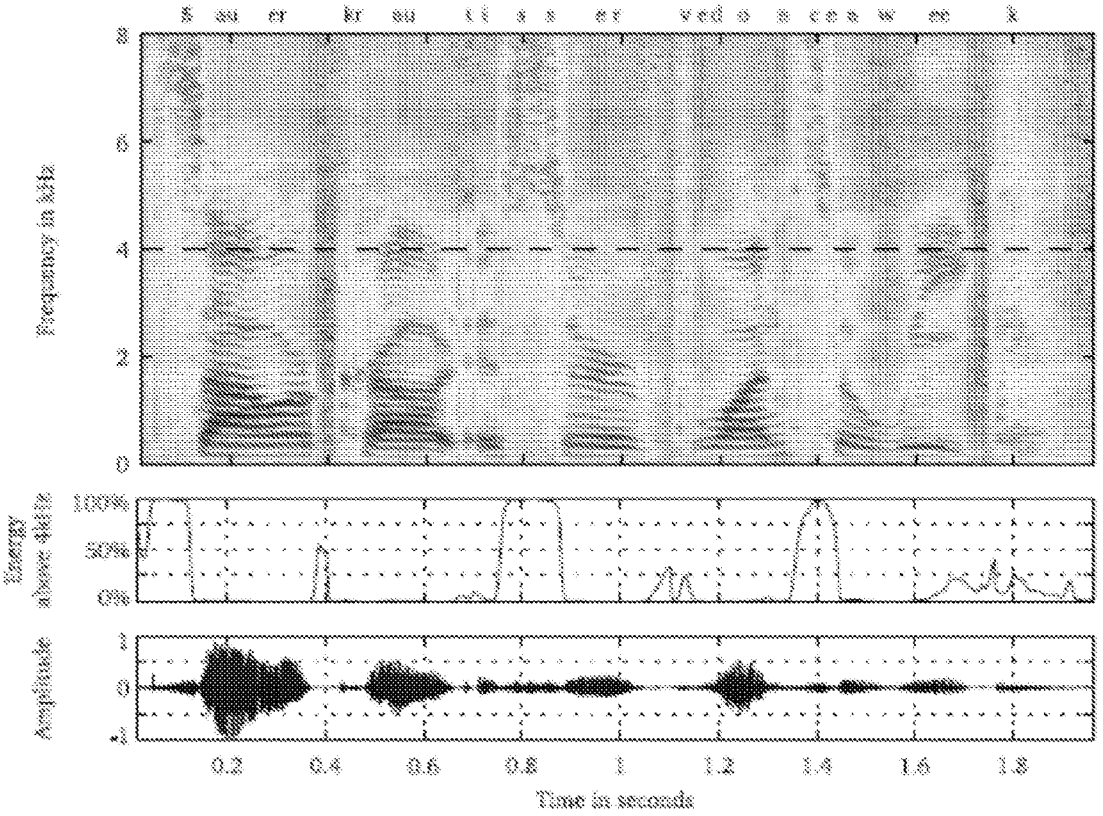

ß au er kr au t i s s er v e d o n ee s w ee k
Frequency in kHz
8
6
4
2
0
Energy above 4kHz
100%
50%
0%
Amplitude
1
0
-1
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Time in seconds

Fig. 1

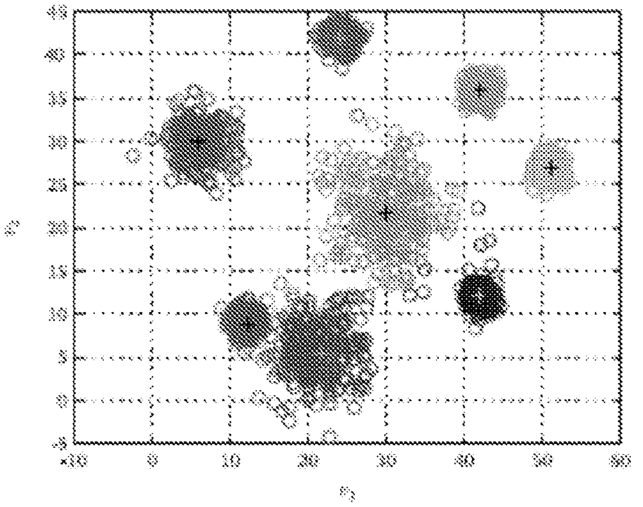

45
40
35
30
25
20
15
10
5
0
-5
c2
-10
0
10
20
30
40
50
60
c1

Fig. 2

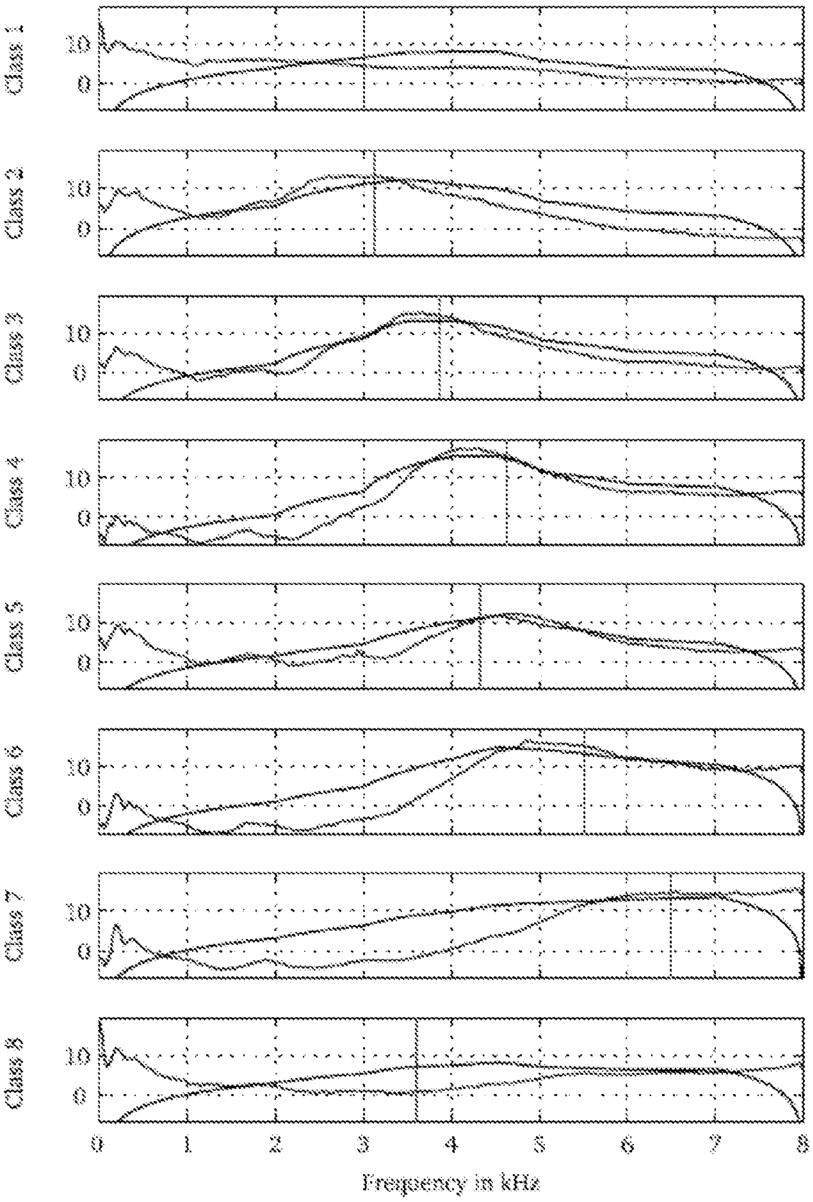

Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
10
0
0
1
2
3
4
5
6
7
8
Frequency in kHz

Fig. 3

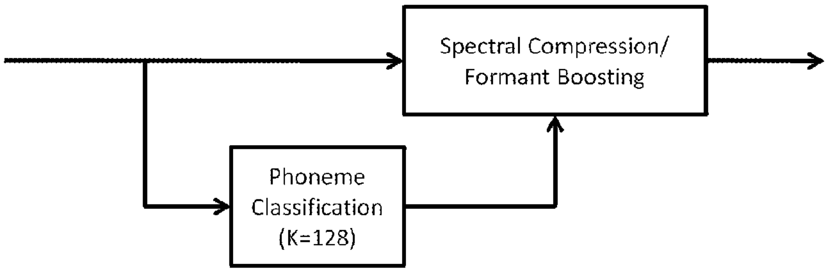


Fig. 4

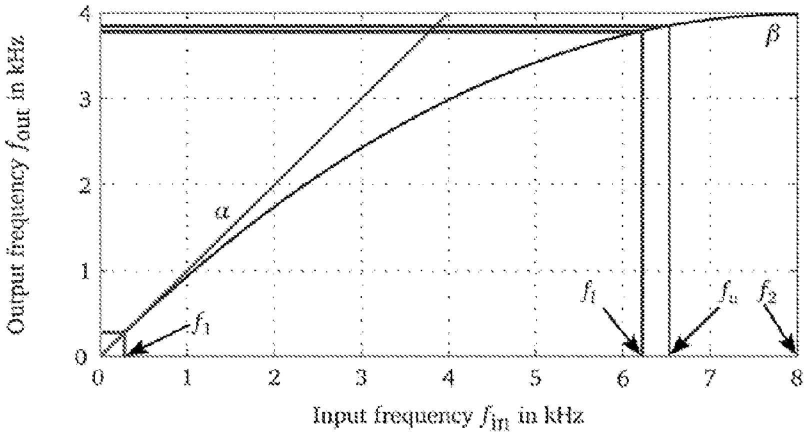


Fig. 5

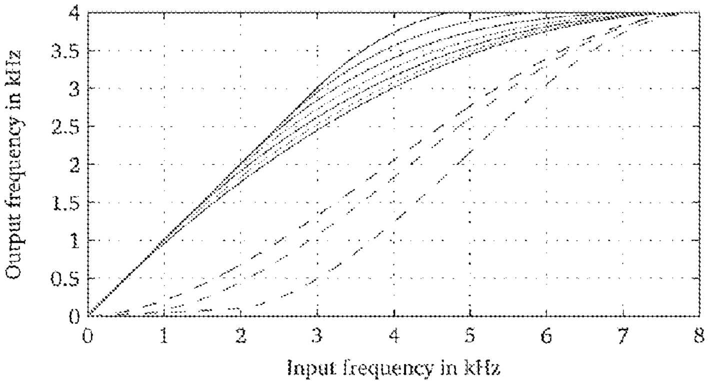


Fig. 6

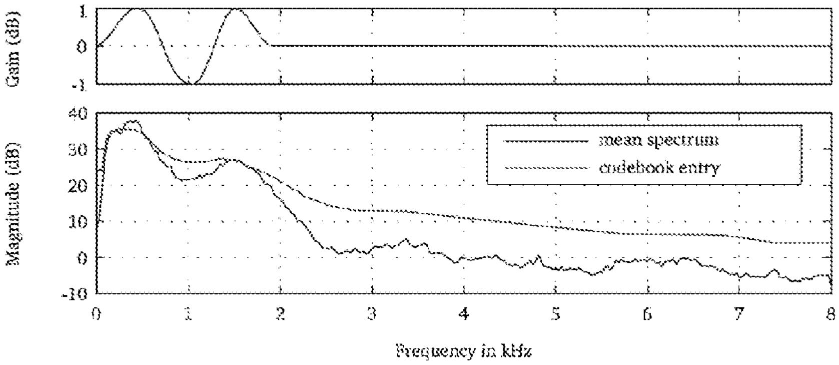


Fig. 7

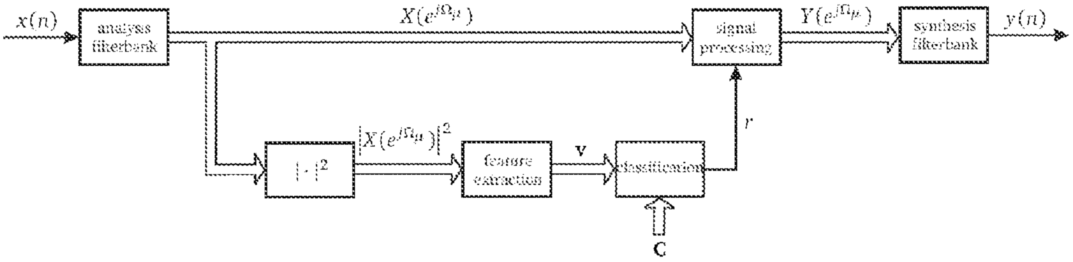
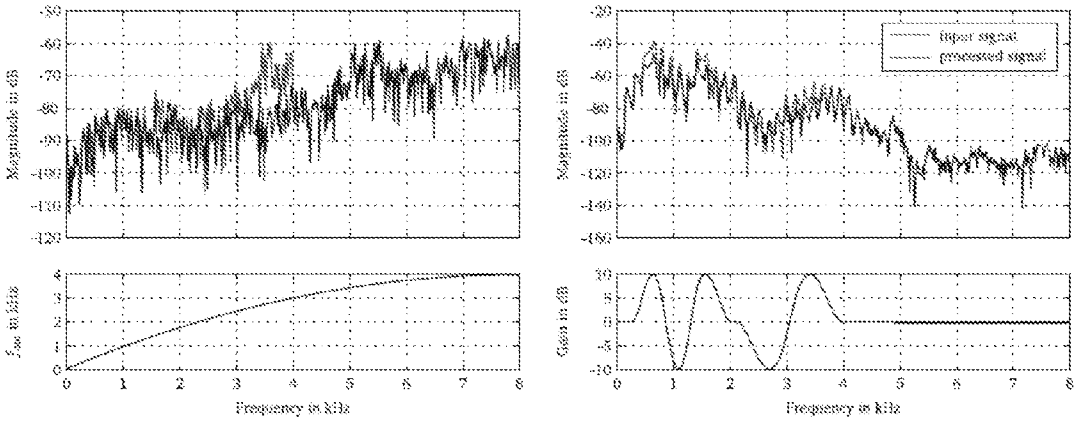


Fig. 8



(a)

(b)

Fig. 9

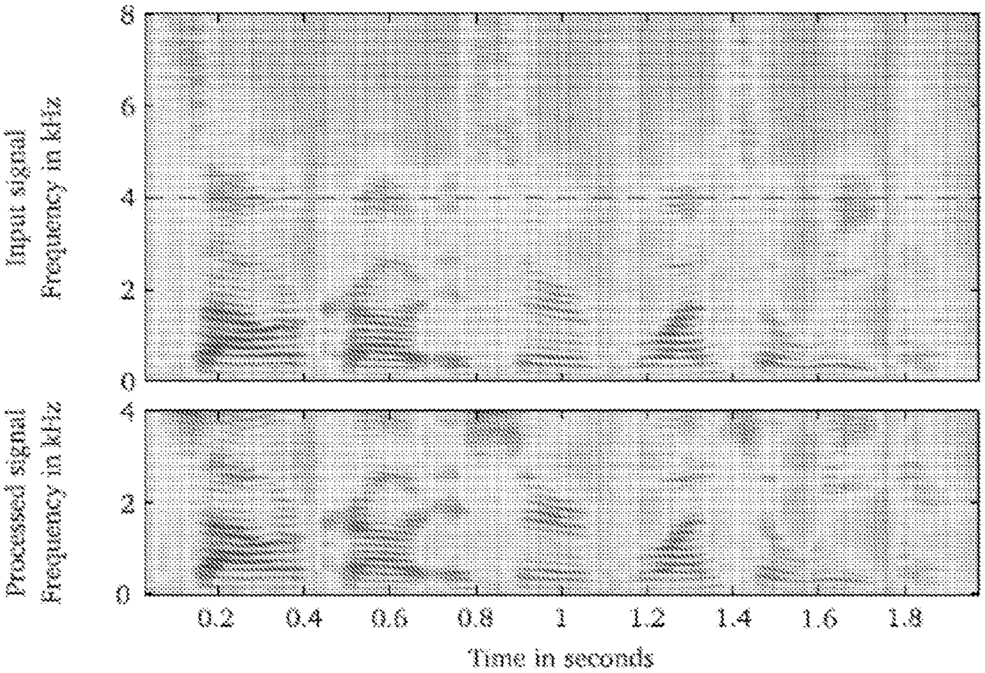

Input signal
Frequency in kHz
0
2
4
6
8
Processed signal
Frequency in kHz
0
2
4
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Time in seconds

Fig. 10

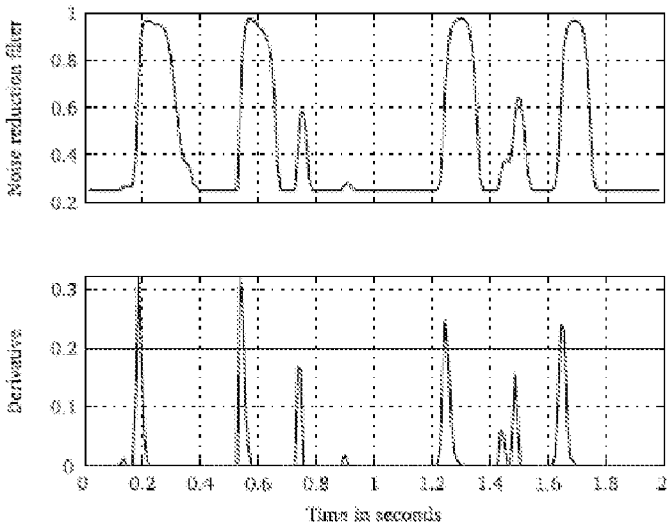

Noise reduction filter
0.2
0.4
0.6
0.8
1
Derivative
0
0.1
0.2
0.3
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Time in seconds

Fig. 11

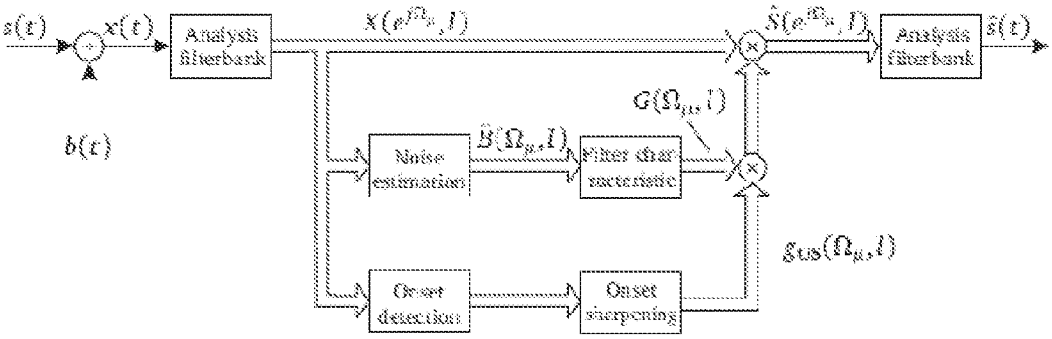

$s(t)$
$x(t)$
Analysis filterbank
$X(e^{j\Omega_\mu}, l)$
$\hat{S}(e^{j\Omega_\mu}, l)$
Analysis filterbank
$\hat{s}(t)$
$b(t)$
$G(\Omega_\mu, l)$
Noise estimation
$\hat{B}(\Omega_\mu, l)$
Filter characteristic
$g_{os}(\Omega_\mu, l)$
Onset detection
Onset sharpening

Fig. 12

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 10 15 6530

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
A	WO 2005/015952 A1 (VAST AUDIO PTY LTD [AU]; CARLILE SIMON [AU]; JIN CRAIG [AU]; LEUNG JOH) 17 February 2005 (2005-02-17) * abstract * * page 8, lines 4-26 * * figures 2,3 * -----	1-15	INV. G10L21/02
A	US 6 577 739 B1 (HURTIG RICHARD RAY [US] ET AL) 10 June 2003 (2003-06-10) * abstract * * column 2, line 40 - column 3, line 9 * * figure 1 * -----	1,15	
A	EP 1 333 700 A2 (PHONAK AG [CH]) 6 August 2003 (2003-08-06) * abstract * * column 6, paragraph [0023] * * column 7, paragraph [0029] * * figures 2-4 * -----	1,15	
			TECHNICAL FIELDS SEARCHED (IPC)
			G10L H04R

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	29 July 2010	Greiser, Norbert

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons
& : member of the same patent family, corresponding document

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 10 15 6530

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

29-07-2010

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2005015952 A1	17-02-2005	AT 452513 T	15-01-2010
		CA 2534139 A1	17-02-2005
		CN 1836465 A	20-09-2006
		DK 1661434 T3	19-04-2010
		EP 1661434 A1	31-05-2006
		ES 2336331 T3	12-04-2010
		NZ 544835 A	31-07-2008
		US 2007127748 A1	07-06-2007
US 6577739 B1	10-06-2003	NONE	
EP 1333700 A2	06-08-2003	DK 1441562 T3	19-07-2010

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 5771299 A **[0014]**
- US 20090074197 A **[0016]**
- EP 1333700 A2 **[0017]**
- US 20060226016 A **[0018]**
- US 20090226016 A **[0019]**
- US 6577739 B **[0020]**
- CA 2569221 **[0021]**
- WO 20080123886 A **[0022]**

Non-patent literature cited in the description

- **P. PATRICK ; R. STEELE ; C. XYDEAS.** *Frequency compression of 7.6 kHz speech into 3.3 kHz bandwidth,* May 1983, vol. 31 (5), 692-701 **[0004]**
- **D. A. HEIDE ; G. S. KANG.** Speech enhancement for bandlimited speech. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,* 12 May 1998, vol. 1, 393-396 **[0007]**
- **ANDREA SIMPSON ; ADAM A. HERSBACH ; HUGH J. MCDERMOTT.** Improvements in speech perception with an experimental nonlinear frequency compression hearing device. *International Journal of Audiology,* 2006, vol. 44 (5), 281-292 **[0011]**
- **H. J. MCDERMOTT ; V. P. DORKOS ; M. R. DEAN ; T. Y. CHING.** Improvements in speech perception with use of the avr transonic frequency-transposing hearing aid. *J Speech Hear Lang Res,* 1999, vol. 42 (6), 1323-1335 **[0012]**
- **KUK ; KORHONEN ; PEETERS ; JESSEN ; ANDERSEN.** Linear frequency transposition: Extending the audibility of high frequency information. *The Hearing Review,* 2006 **[0012]**