(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 September 2015 (11.09.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/134579 A1

(51) International Patent Classification:
G10L 15/14 (2006.01)

(21) International Application Number:
PCT/US2015/018692

(22) International Filing Date:
4 March 2015 (04.03.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/947,841 4 March 2014 (04.03.2014) US

(71) Applicant: **INTERACTIVE INTELLIGENCE GROUP, INC.** [US/US]; 7601 Interactive Way, Indianapolis, Indiana 46278 (US).

(72) Inventors: **CHELUVARAJA, Srinath**; Interactive Intelligence, Inc., 7601 Interactive Way, Indianapolis, Indiana 46278 (US). **IYER, Ananth Nagaraja**; Interactive Intelligence, Inc., 7601 Interactive Way, Indianapolis, Indiana 46278 (US). **GANAPATHIRAJU, Aravind**; #46 Bhavya's Alluri Meadows, Whitefields, Kondapur, Hyderabad, AP (IN). **WYSS, Felix Immanuel**; Interactive Intelligence, Inc., 7601 Interactive Way, Indianapolis, Indiana 46278 (US).

(74) Agents: **COLE, Troy, J.** et al.; Ice Miller, LLP, One American Square, Suite 2900, Indianapolis, Indiana 46282 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

[Continued on next page]

(54) Title: SYSTEM AND METHOD TO CORRECT FOR PACKET LOSS IN ASR SYSTEMS

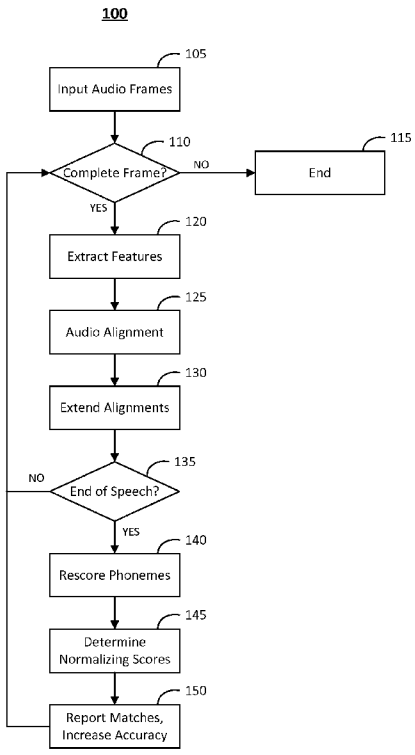

Fig. 1

(57) Abstract: A system and method are presented for the correction of packet loss in audio in automatic speech recognition (ASR) systems. Packet loss correction, as presented herein, occurs at the recognition stage without modifying any of the acoustic models generated during training. The behavior of the ASR engine in the absence of packet loss is thus not altered. To accomplish this, the actual input signal may be rectified, the recognition scores may be normalized to account for signal errors, and a best-estimate method using information from previous frames and acoustic models may be used to replace the noisy signal.

WO 2015/134579 A1

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

SYSTEM AND METHOD TO CORRECT FOR PACKET LOSS IN ASR SYSTEMS

BACKGROUND

[1] The present invention generally relates to telecommunications systems and methods, as well as automatic speech recognition systems. More particularly, the present invention pertains to the correction of packet loss within the systems.

SUMMARY

[2] A system and method are presented for the correction of packet loss in audio in automatic speech recognition (ASR) systems. Packet loss correction, as presented herein, occurs at the recognition stage without modifying any of the acoustic models generated during training. The behavior of the ASR engine in the absence of packet loss is thus not altered. To accomplish this, the actual input signal may be rectified, the recognition scores may be normalized to account for signal errors, and a best-estimate method using information from previous frames and acoustic models may be used to replace the noisy signal.

[3] In one embodiment, a method to correct for packet loss in an audio signal in an automatic speech recognition system comprising reconstructing the signal using linear interpolation from preceding and succeeding frames of audio is presented, wherein the linear interpolation comprises the steps of: determining spectral vectors from a time interval of frames prior to a packet loss event; maintaining a sliding buffer, comprising a number of frames, between the time interval of frames and the packet loss event; generating the linear interpolation for the time interval of frames; using energy of the signal to tag frames within the packet loss event; and inserting values for the packet loss event into the sliding buffer.

[4] In another embodiment, a method to correct for packet loss in an audio signal in automatic speech recognition systems comprising rescoring of phoneme probability calculations for a section of the audio signal is presented, comprising the steps of: accumulating probabilities for a given word in a lexicon; determining when the probabilities are statistically significant; reporting matches when the probabilities

are statistically significant; normalizing values of probabilities that are not statistically significant; and rescoring the probabilities for the given word.

[5] In another embodiment, A method to correct for packet loss in an audio signal in automatic speech recognition systems comprising maximum likelihood prediction wherein the maximum likelihood prediction further comprises the steps of: determining Hidden Markov Model state occupation probabilities in an acoustic model within the automatic speech recognition system prior to a packet loss event; and predicting feature values of packets in the packet loss section of the audio signal based on the determined Hidden Markov Model state occupation probabilities.

BRIEF DESCRIPTION OF THE DRAWINGS

[6] Figure 1 is a flowchart illustrating an embodiment of a process for phoneme rescoring.

[7] Figure 2 is a flowchart illustrating an embodiment of a process for determining maximum likelihood.

[8] Figure 3 is a flowchart illustrating an embodiment of a process for determining interpolation.

DETAILED DESCRIPTION

[9] For the purposes of promoting an understanding of the principles of the invention, reference will now be made to the embodiment illustrated in the drawings and specific language will be used to describe the same. It will nevertheless be understood that no limitation of the scope of the invention is thereby intended. Any alterations and further modifications in the described embodiments, and any further applications of the principles of the invention as described herein are contemplated as would normally occur to one skilled in the art to which the invention relates.

[10] ASR systems are typically used for recognizing input audio from users for extracting relevant information from the input audio. The relevant information may include digits of a spoken telephone number, keywords in some predefined domain, or, in the case of larger recognition systems, even more complex phrases and sentences. ASR systems use hidden Markov (HMM) models whose parameters are trained from a large corpus of known audio data from a specified domain. The recognition of user audio

may be made by matching the audio to existing models using efficient methods to streamline the search and cull the number of candidates.

[11] ASR systems generally comprise training units and recognition units. During training, lower dimensional (e.g., 39) feature data is extracted from each audio frame in a large corpus and HMMs are trained for each language unit using this data. Feature data is obtained by first transforming the audio signal into the frequency domain using fast Fourier transform (FFT) and performing further transformations. The extracted features are used to train acoustic models that have the same distribution as the training data. Methods from statistics, such as expectation maximization, may be used. During recognition, the extracted features from input audio are matched to the model that maximizes the likelihood of the given feature data. The maximization process is continued for every ensuing frame of audio using an efficient dynamic programming determination. When a statistically significant score is obtained for a word or phrase, a result is reported.

[12] The real-time performance of an ASR system depends on the quality of the models used in training and on the quality of the input audio to be recognized. Audio with large distortions or errors, such as packet loss, results in sharply lower accuracy and degradation in the overall experience. Packet loss, in an embodiment, may be defined as data frames containing zeroes instead of actual transmitted values, in an ASR system.

[13] In VoIP systems, where audio is treated as network data, all traffic occurs in the form of packets. Packet loss is a common source of error caused by router buffer overflows during packet transmission. Correction of packet loss is important for the robustness of the system.

[14] Audio with packet loss or noise will produce poor matches, even to well-trained models, resulting in lower accuracy and high false alarm rates. The embodiments described herein address this problem at the recognition state of the ASR system, without modifying any of the acoustic models generated during training. Prescriptions are provided at different levels: (1) rectification of the actual input signal that is either noisy or consisting of digital silence due to dropped packets, (2) correction in the recognition step

where only the recognition scores are normalized to account for signal errors, and (3) replacement of the noisy signal by a best-estimate using information from previous frames and acoustic models.

[15] Embodiments discussed within can be accommodated in ASR systems at two different stages of the ASR engine: processing of the signal obtained prior to the ASR step or directly at the recognition step without any processing of the faulty input signal.

[16] Figure 1 is a flowchart illustrating an embodiment of a process for phoneme rescoring. The distorted or missing signal is not corrected (except for packet repetition when there is digital silence) but fed directly into the next stage of the ASR engine. The process in Figure 1 is instead applied towards the very end of the ASR step when phoneme probabilities are calculated and being reported for a section of audio.

[17] In operation 105, audio frames are input. For example, the input comprises audio from users that will be recognized by the ASR system for the extraction of relevant information. Control is passed to operation 110 and the process 100 continues.

[18] In operation 110, it is determined whether or not there is a complete frame in the sequence of input audio frames. If it is determined that there is a complete frame in the sequence of input audio frames, control is passed to operation 120 and the process continues. If it is determined that there is not a complete frame in the sequence of input audio frames, control is passed to operation 115 and the process ends.

[19] The determination in operation 110 may be based on any suitable criteria. In an embodiment, the input audio is segmented into overlapping frames. The frames may be 20 ms in length. Segmentation is performed using a Hamming window and analysis is performed only when a complete frame of audio containing at least 20 ms of data is available.

[20] In operation 120, me l frequency cepstral coefficient features are extracted and packet loss is tagged. Packet repetition may be used to calculate features for digital zeros. For example, feature data may be obtained by first transforming the audio signal into the frequency domain using FFT and then performing further transformations using filter banks. In an embodiment, the audio frames used for

feature extraction are decomposed into overlapping frames of size 20 ms with an overlap factor of 50%. Control is passed to operation 125 and the process 100 continues.

[21] In operation 125, the audio is aligned. For example, the extracted features from the input audio is matched to the acoustic model that maximizes the likelihood of the given feature data. The tagged packet information may be utilized for the audio alignment. Control is passed to operation 130 and the process 100 continues.

[22] In operation 130, alignments are extended. In an embodiment, the alignments obtained previously are extended using data from the current frame. As every alignment has a probability associated with it, a list of alignments is maintained and constantly extended and pruned with only the best alignments retained at every step. Control is passed to operation 135 and the process 100 continues.

[23] In operation 135, it is determined whether or not speech has ended. If it is determined that speech has not ended, control is passed back to operation 110 and process 100 continues. If it is determined that speech has ended, control is passed to operation 140 and process 100 continues.

[24] The determination in operation 135 may be made based on any suitable criteria. In an embodiment, a voice activity detector may be used for determining whether or not speech has ended.

[25] In operation 140, phonemes are rescored. In an embodiment, previously obtained phoneme scores from a large corpus may be used in the rescoring step by comparing them to the existing phoneme scores. The rescoring strategy simply tags those words or phrases whose confidence values through their scores has been affected by packet loss using tagged packet information from operation 120. Control is passed to operation 145 and process 100 continues.

[26] In operation 145, normalized word scores are determined. In an embodiment, two normalization techniques are used: (1) Phoneme scores are deleted for sections with packet loss and word scores are recalculated excluding dropped sections, and (2) Phoneme scores for lossy sections are replaced with historical values obtained offline and recalculated. Control is passed to operation 150 and process 100 continues.

[27] In operation 150, matches are reported and the accuracy of the system is increased. In an embodiment, the normalized scores from operation 145 are converted to confidence values and matches are reported when confidence exceeds a pre-defined threshold. Control is passed back to operation 110 and process 100 continues.

[28] During normal operation without packet loss, probabilities for a given word in the lexicon are accumulated using the highest scoring tokens for every frame of audio and matches are reported when these scores become statistically significant. Low confidence hits are not reported and count as misses and reduction in false alarms. The effect of packet loss is to reduce the confidence values of many possible matches because of low scores obtained by mismatching faulty audio in certain sections of otherwise high confidence hits. This is especially true when packet loss extends over a short section of a phrase or sentence.

[29] The historical calculation needs the identity of the phoneme that is being normalized and this is possible when a long phoneme has packet loss in a short section. Historical values are previously obtained for each phoneme from a corpus without any packet loss. The effect of both these steps is a big improvement in overall ASR recognition accuracy (2 to 4%) but also triggers additional false alarms. The false alarms are controlled by applying either of the above two normalization methods only to likely candidates and several heuristics are used like value of the token scores, value of token score shifts from previous values, etc.

[30] Figure 2 is a flowchart illustrating an embodiment of a process for determining maximum likelihood. The HMM state occupation probabilities in the acoustic model just prior to a packet loss event are known and the most likely feature values of the missing packet are predicted conditionally on these values.

[31] In operation 205, audio frames are input. For example, the input comprises audio from users that will be recognized by the ASR system for the extraction of relevant information. Control is passed to operation 210 and the process 200 continues.

[32] In operation 210, it is determined whether or not there is a complete frame in the sequence of input audio frames. If it is determined that there is a complete frame in the sequence of input audio frames, control is passed to operation 220 and the process continues. If it is determined that there is not a complete frame in the sequence of input audio frames, control is passed to operation 215 and the process ends.

[33] The determination in operation 210 may be based on any suitable criteria. In an embodiment, the input audio is segmented into overlapping frames. The frames may be 20 ms in length with some overlapping factor, such as 50%. Segmentation is performed using a Hamming window and analysis is performed only when a complete frame of audio containing at least 20 ms of data is available.

[34] In operation 220, it is determined whether or not the frame is a lossy packet. If it is determined that the frame is a lossy packet, control is passed to operation 225. If it is determined that the frame is not a lossy packet, control is passed to operation 230.

[35] The determination in operation 220 may be based on any suitable criteria. In an embodiment, a voice activity detector may be used for determining packet loss results. Packet loss results in digital zeros in the audio frame.

[36] In operation 225,features are predicted. In an embodiment, the most likely feature values of the succeeding packet are found by maximizing a probability function that is a sum over several Gaussian mixtures with weights given by the state occupation probabilities. A previously trained acoustic model may be utilized in the prediction.

[37] Feature prediction may be determined as follows. Let x(1), x(2) … x(t) represent feature vectors observed until time t just before packet loss at t+1. The probability distribution of the feature vector at x(t+1) may be represented mathematically by:

[38] $P\big(x(t+1)\big|x(1),x(2)\dots x(t)\big) = \sum_{i\,k} alpha(i)t(ik)b(k)\big(x(t+1)\big)$

[39] where $alpha(i)$ represents the state occupation probabilities at time t for state (i) (this may also be referred to as the forward variable), where $t(ik)$ represents the state transition probabilities from state i

to k, and where $b(k)(x(t+1))$ represents the Gaussian mixture distribution for the feature x(t+1) from a state k belonging to a phoneme model.

[40] The best estimate for x(t+1) is given by maximizing $P(x(t+1)|x(1), x(2) \ldots x(t))$ over all possible x(t+1). The full maximizing of this function would introduce significant overheads and instead, approximations are used to predict these feature values. Control is passed to operation 235 and the process 200 continues.

[41] In operation 230, me l frequency cepstral coefficient features are extracted. For example, feature data may be obtained by first transforming the audio signal into the frequency domain using FFT and then performing further transformations using filter banks. In an embodiment, the audio frames used for feature extraction are decomposed into overlapping frames of size 20 ms with an overlap factor. Control is passed to operation 235 and process 200 continues.

[42] In operation 235, the audio is aligned. For example, the extracted features from the input audio is matched to the acoustic model that maximizes the likelihood of the given feature data. The tagged packet information may be utilized for the audio alignment. Control is passed to operation 240 and process 200 continues.

[43] In operation 240, alignments are extended. In an embodiment, the alignments obtained previously are extended using data from the current frame. As every alignment has a probability associated with it, a list of alignments is maintained and constantly extended and pruned with only the best alignments retained at every step. Control is passed to operation 245 and process 200 continues.

[44] In operation 245, the current state information is recorded. The current state information comprises the most recent state of the forward variable of all the audio alignments and needs to be available as packet loss events are unpredictable. Control is passed to operation 250 and process 200 continues.

[45] In operation 250, it is determined if speech has ended. If it is determined that speech has ended, control is passed to operation 255 and process 200 continues. If it is determined that speech has not ended, control is passed back to operation 210 and the process 200 continues.

[46] The determination in operation 250 may be based on any suitable criteria. In an embodiment, a voice activity detector may be used for determining whether or not speech has ended.

[47] In operation 255, the confidence is determined. Control is passed to operation 260 and the process 200 continues.

[48] In operation 260, matches are reported and the accuracy of the system is increased. In an embodiment, the normalized scores in 255 are converted to confidence values and matches are reported when confidence exceeds a pre-defined threshold. Control is passed back to operation 210 and the process 200 continues.

[49] Figure 3 is a flowchart illustrating an embodiment of a process for determining interpolation. Speech waveforms are reasonably modeled as the output of linear systems with slowly varying filter parameters. The missing, or distorted, signal is reconstructed using linear interpolation from preceding and succeeding audio frames. Spectral data in the neighboring frames are used to generate linear parameters which are then used to fill in the values in the distorted frames.

[50] In operation 305, audio frames are input. For example, the input comprises audio from users that will be recognized by the ASR system for the extraction of relevant information. Control is passed to operation 310 and the process 300 continues.

[51] In operation 310, it is determined whether or not there is a complete frame in the sequence of input audio frames. If it is determined that there is a complete frame in the sequence of input audio frames, control is passed to operation 320 and the process continues. If it is determined that there is not a complete frame in the sequence of input audio frames, control is passed to operation 315 and the process ends.

[52] The determination in operation 310 may be based on any suitable criteria. In an embodiment, the input audio is segmented into overlapping frames. The frames may be 20 ms in length. Segmentation is

performed using a Hamming window and analysis is performed only when a complete frame of audio containing at least 20 ms of data is available.

[53] In operation 320, lossy packets are tagged for the following interpolation operations. In an embodiment, this may be done by checking for digital zeros or zero energy. Control is passed to operation 325 and process 300 continues.

[54] In operation 325, the interpolation buffer is updated. A sliding buffer may be used to correct for packet loss. In an embodiment, a sliding buffer of 10 frames (k=4) to correct for possible packet loss in the central 2 frames is used. The signal energy is then used to tag frames with packet loss and these are differentiated from ordinary silence. In the case of burst losses (defined as more than two successive frames with digital silence) many elements of the buffer are zeros and packet repetition is used to fill in the values in the sliding buffer for the interpolation step. As the buffer slides with more incoming data, corrected values are used to replace lossy values. Control is passed to operation 330 and process 300 continues.

[55] In operation 330, it is determined whether or not the buffer is full. If it is determined that the buffer is full, control is passed to operation 335 and the process 300 continues. If it is determined that the buffer is not full, control is passed back to operation 310 and the process 300 continues.

[56] The determination in operation 330 may be made based on any suitable criteria. In an embodiment, the buffer is full after the first 10 frames are accommodated by discarding the oldest members.

[57] In operation 335, central packets are interpolated. For example, let x(1), x(2) … x(t) represent spectral vectors (256 dimensional) until time t before the first packet loss event occurs at time t+1. A buffer is maintained that contains x(t-k+1) … x(t), x(t+3), x(t+4) … x(t+2+k), 2K +2 frames in all, where k is a small number (e.g. 4). The values of x(t-k+1) … x(t), x(t+3) … x(t+3+k) are used (2k in all) to generate a linear interpolation for x(t+1) and x(t+2). In an embodiment, two frames are simultaneously interpolated to keep in step with the keyword spotter.

[58] The interpolation step is optimal when no more than 2 to 3 frames are consecutively zero because there is then enough clean data in the buffer to generate interpolation parameters. The average interpolation error is used to monitor the quality of the interpolation and to decide if it needs to be applied at all. Assuming that the first k frames have no packet loss, the sliding buffer will always have corrected frames in the first half. Control is passed to operation 340 and the process 300 continues.

[59] In operation 340, me l frequency cepstral coefficient features are extracted. For example, feature data may be obtained by first transforming the audio signal into the frequency domain using FFT and then performing further transformations using filter banks. In an embodiment, the audio frames used for feature extraction are decomposed into overlapping frames of size 20 ms with an overlap factor. Control is passed to operation 345 and the process 300 continues.

[60] In operation 345, the audio is aligned. For example, the extracted features from the input audio is matched to the acoustic model that maximizes the likelihood of the given feature data. Control is passed to operation 350 and the process 300 continues.

[61] In operation 350, it is determined whether or not the speech has ended. If it is determined that the speech has ended, control is passed to operation 355 and the process 300 continues. If it is determined that the speech has not ended, control is passed back to operation 310 and process 300 continues.

[62] The determination in operation 350 may be made based on any suitable criteria. In an embodiment, a voice activity detector may be used for determining whether or not speech has ended.

[63] In operation 355, the confidence value is determined. Control is passed to operation 360 and process 300 continues.

[64] In operation 360, matches are reported and the accuracy of the system is increased. In an embodiment, the normalized scores in 355 are converted to confidence values and matches are reported when confidence exceeds a pre-defined threshold. Control is passed back to operation 310 and the process 300 continues.

[65] For larger bursts, the interpolation parameters become unreliable and do not yield any improvement over packet repetition. In an embodiment, these parameters may be generated by a

minimum least squares approach assuming a linear or quadratic fit. The new spectral parameters are then fed into the next stages of the ASR engine for feature calculation, etc., instead of the old values where packet repetition is used to correct for packet loss. No other additional operations are performed downstream in the ASR engine. The main feature of this approach is directly operating on the spectrum instead of the features. More buffer storage space and processing is needed to handle large dimensional vectors. The effect of this method is a marginal but non-zero improvement in accuracy (0.3%) and no additional false alarms are generated.

[66] While the invention has been illustrated and described in detail in the drawings and foregoing description, the same is to be considered as illustrative and not restrictive in character, it being understood that only the preferred embodiment has been shown and described and that all equivalents, changes, and modifications that come within the spirit of the invention as described herein and/or by the following claims are desired to be protected.

[67] Hence, the proper scope of the present invention should be determined only by the broadest interpretation of the appended claims so as to encompass all such modifications as well as all relationships equivalent to those illustrated in the drawings and described in the specification.

CLAIMS

1. A method to correct for packet loss in an audio signal in an automatic speech recognition system comprising reconstructing the signal using linear interpolation from preceding and succeeding frames of audio, wherein the linear interpolation comprises the steps of:
    a. determining spectral vectors from a time interval of frames prior to a packet loss event;
    b. maintaining a sliding buffer, comprising a number of frames, between the time interval of frames and the packet loss event;
    c. generating the linear interpolation for the time interval of frames;
    d. using energy of the signal to tag frames within the packet loss event; and
    e. inserting values for the packet loss event into the sliding buffer.
2. The method of claim 1, wherein at least two frames undergo linear interpolation simultaneously.
3. The method of claim 1, wherein the sliding buffer comprises 10 frames.
4. The method of claim 1, wherein a packet loss event comprises digital silence.
5. The method of claim 1, wherein a packet loss event comprises more than two successive frames of digital silence.
6. The method of claim 5, wherein values are inserted into the more than two successive frames of digital silence using packet repetition.
7. The method of claim 1, wherein a packet loss event comprises a large number of successive frames of digital silence.
8. The method of claim 7, wherein values are inserted into the large number of successive frames with digital silence using a minimum least squares approach.
9. The method of claim 1, wherein the time interval of frames are decomposed into overlapping frames of 20 ms with an overlap factor.
10. A method to correct for packet loss in an audio signal in automatic speech recognition systems comprising rescoring of phoneme probability calculations for a section of the audio signal, comprising the steps of:

a. accumulating probabilities for a given word in a lexicon;

b. determining when the probabilities are statistically significant;

c. reporting matches when the probabilities are statistically significant;

d. normalizing values of probabilities that are not statistically significant; and

e. rescoring the probabilities for the given word.

11. The method of claim 10 wherein the normalizing of step (d) comprises deleting values for sections of the audio signal comprising packet loss.

12. The method of claim 10 wherein the normalizing of step (d) comprises replacing values for sections of the audio signal comprising packet loss with historical values obtained offline.

13. The method of claim 10, wherein the audio frames are decomposed into overlapping frames of 20 ms with an overlap factor.

14. A method to correct for packet loss in an audio signal in automatic speech recognition systems comprising maximum likelihood prediction wherein the maximum likelihood prediction further comprises the steps of:

a. determining Hidden Markov Model state occupation probabilities in an acoustic model within the automatic speech recognition system prior to a packet loss event; and

b. predicting feature values of packets in the packet loss section of the audio signal based on the determined Hidden Markov Model state occupation probabilities.

15. The method of claim 14, wherein the Hidden Markov Model state occupation probabilities comprise a sum over several Gaussian mixtures with weights given by the state occupation probabilities.

16. The method of claim 15, wherein the state occupation probabilities are determined using the mathematical equation:

$$P\big(x(t+1)\big|x(1),x(2)\dots x(t)\big) = \sum_{i\,k} alpha(i)t(ik)b(k)(x(t+1))$$

17. The method of claim 14, wherein the prediction of the feature values further comprises determining if a feature value is at a first state of the acoustic model and if the next state is at the first state or at a second state.
18. The method of claim 14, wherein the prediction of the feature values further comprises determining if a feature value is at a second state and the next state is in the same state or an other state of the acoustic model.
19. The method of claim 18, wherein the other state of the acoustic model comprises a last state.
20. The method of claim 19, wherein the next state after the last state must be the same state or the first state of the next node in a lexical network.
21. The method of claim 14 wherein the audio frames are decomposed into overlapping frames of 20 ms with an overlap factor.

100

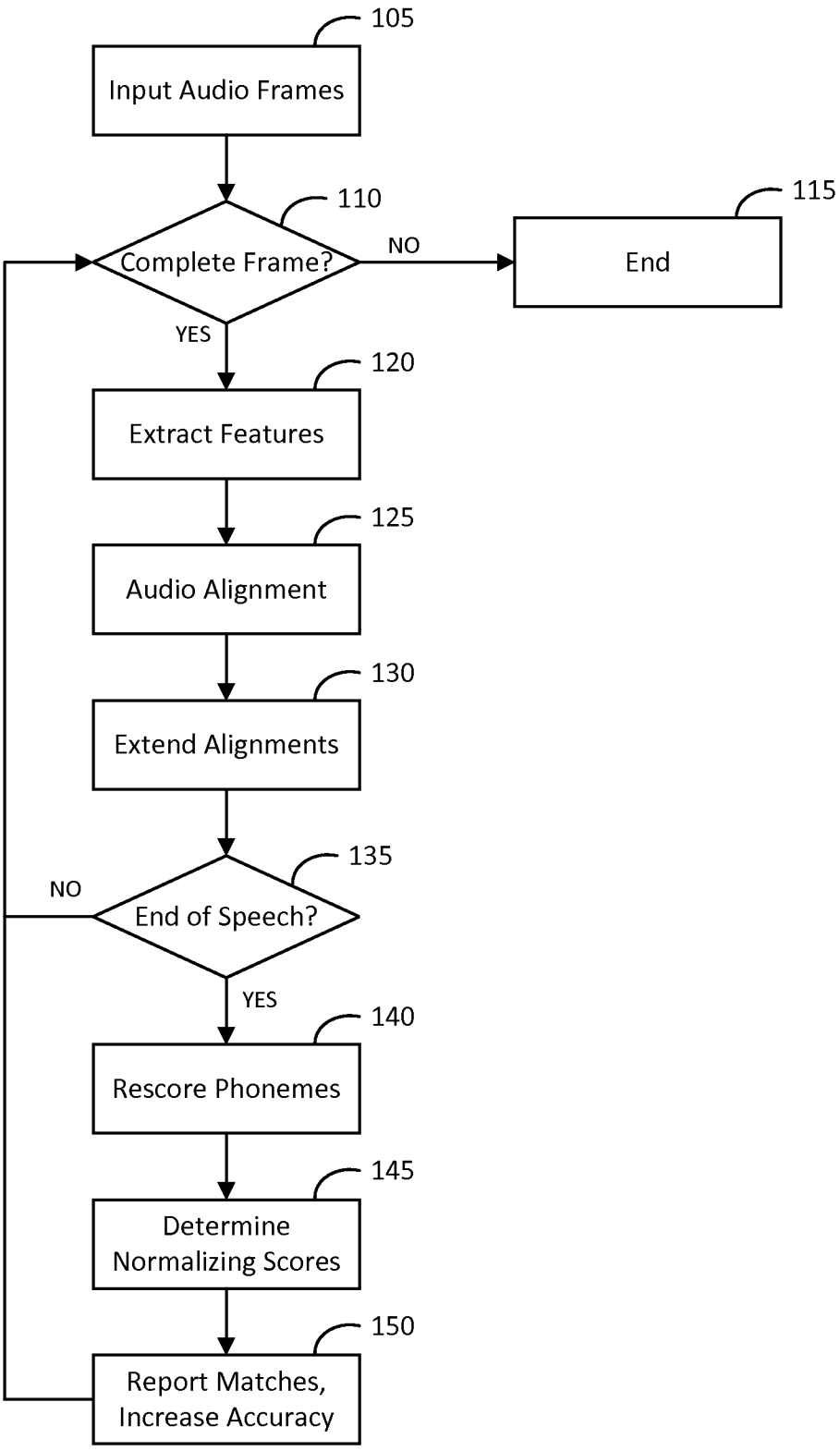
105
Input Audio Frames
110
Complete Frame?
NO
115
End
YES
120
Extract Features
125
Audio Alignment
130
Extend Alignments
135
NO
End of Speech?
YES
140
Rescore Phonemes
145
Determine Normalizing Scores
150
Report Matches, Increase Accuracy

Fig. 1

200

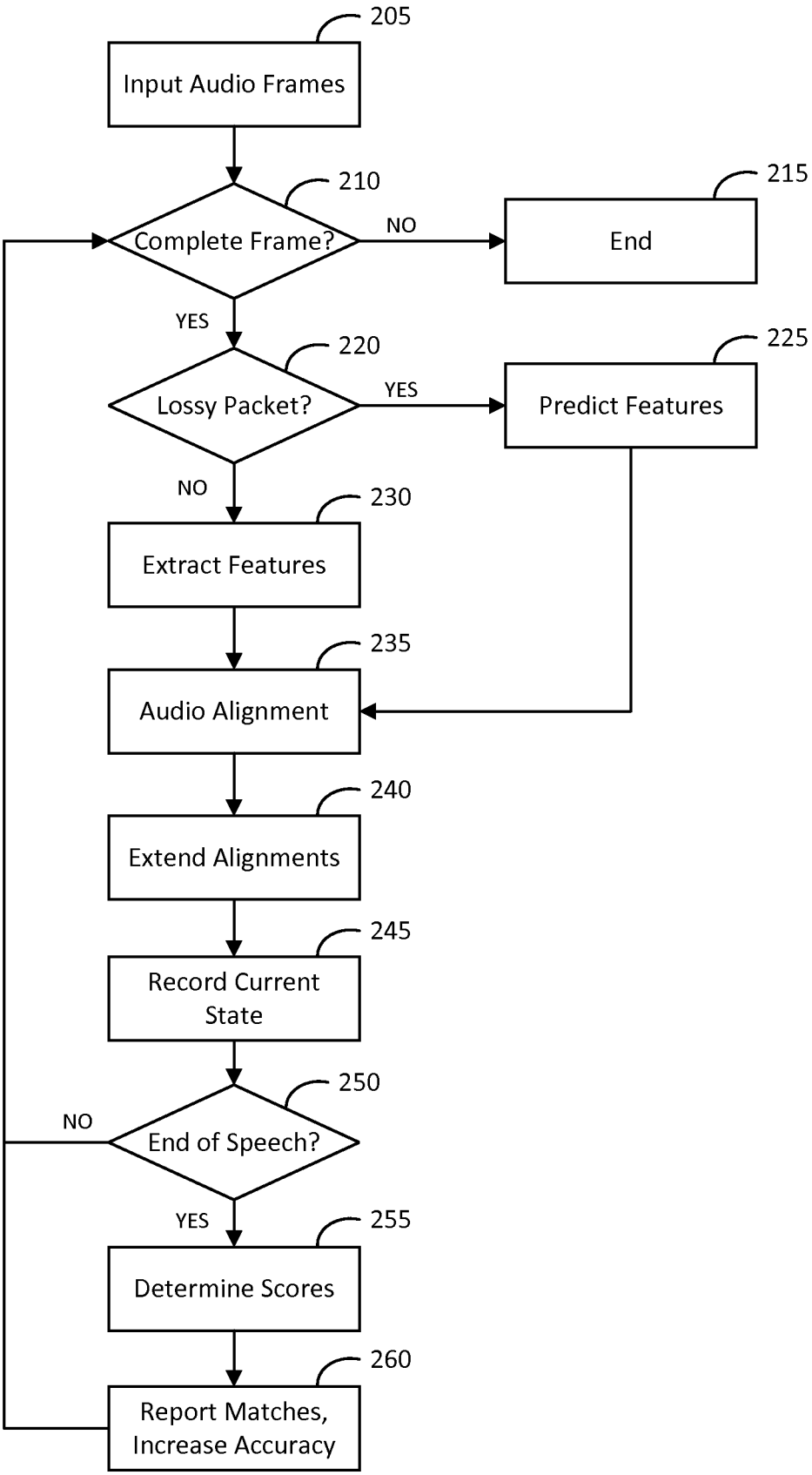
205
Input Audio Frames
210
Complete Frame?
NO
215
End
YES
220
Lossy Packet?
YES
225
Predict Features
NO
230
Extract Features
235
Audio Alignment
240
Extend Alignments
245
Record Current State
250
NO
End of Speech?
YES
255
Determine Scores
260
Report Matches, Increase Accuracy

Fig. 2

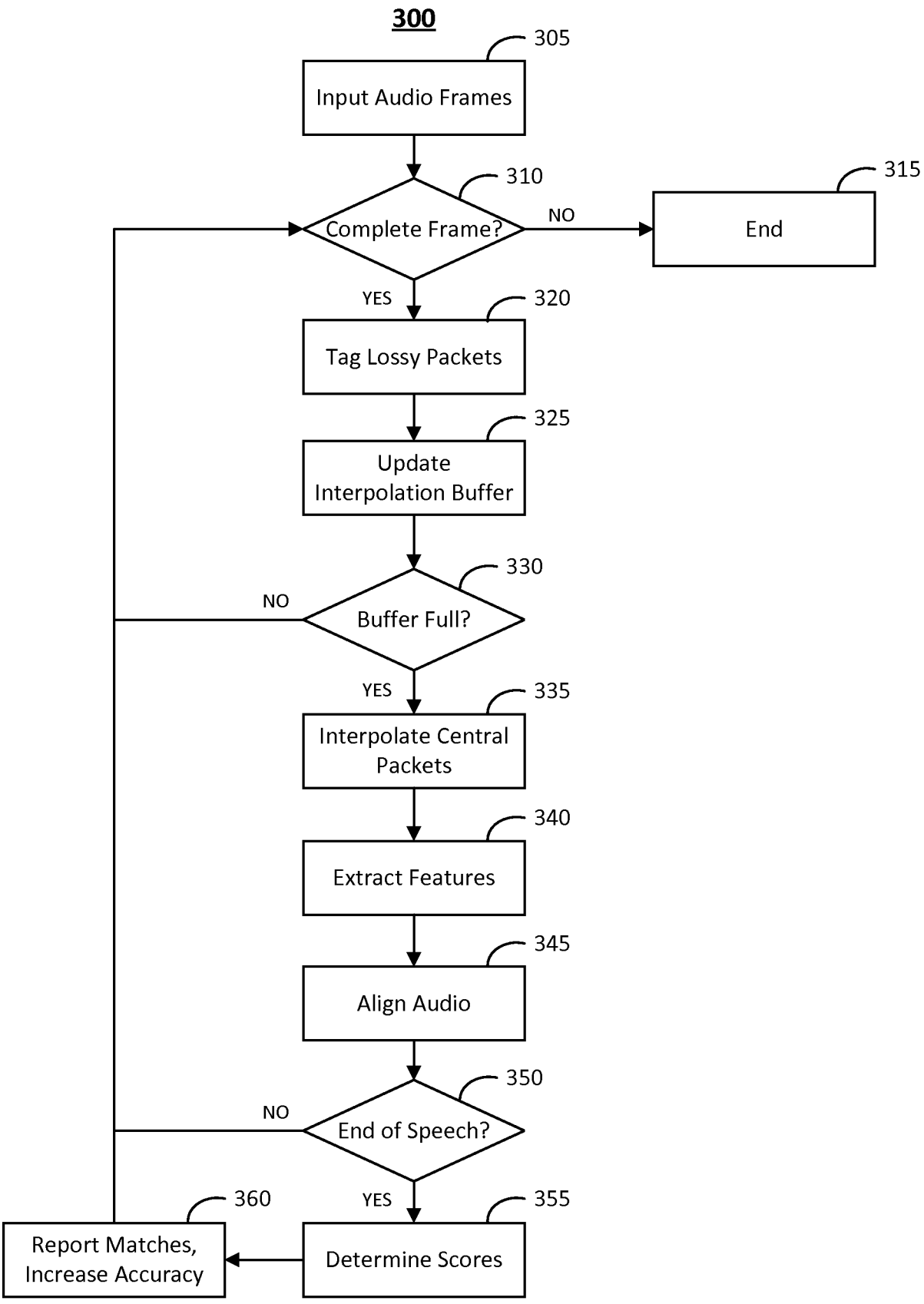
300
305
Input Audio Frames
310
Complete Frame?
NO
315
End
YES
320
Tag Lossy Packets
325
Update Interpolation Buffer
330
Buffer Full?
NO
YES
335
Interpolate Central Packets
340
Extract Features
345
Align Audio
350
End of Speech?
NO
YES
355
Determine Scores
360
Report Matches, Increase Accuracy

Fig. 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US15/18692

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G10L 15/14, 19/04 (2015.01)
CPC - G10L 15/142, 19/06
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) Classification(s): G10L 15/02, 15/14, 15/18, 19/04, 25/18 (2015.01)
CPC Classification(s): G10L 15/02, 15/142, 15/18, 19/04, 19/06, 25/18

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PatSeer (US, EP, WO, JP, DE, GB, CN, FR, KR, ES, AU, IN, CA, INPADOC Data)
Keywords: ASR, speech recognition, audio, voice, acoustic, packet, frame, loss, erase, correct, conceal, normalize, HMM, Markov, Gaussian, buffer, time interval, signal, energy, feature, linear interpolation

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y --- A	US 2010/0174537 A1 (VOS, K. et al.) 08 July 2010; paragraphs [0085], [0092], [0114], [0116], [0151].	1-2 --- 3-9 --- 16
X --- Y	US 2012/0116766 A1 (WASSERBLAT, M. et al.) 10 May 2012; paragraphs [0056], [0064].	10 --- 11-13
X --- Y	KOENIG et al. "A NEW FEATURE VECTOR FOR HMM-BASED PACKET LOSS CONCEALMENT"; 17th European Signal Processing Conference (EUSIPCO 2009); Publication [online]. 28 August 2009 [retrieved 09 July 2015]. Retrieved from the Internet: <URL: http://www.eurasip.org/Proceedings/Eusipco/Eusipco2009/contents/papers/1569191501.pdf pp 2519-2523.	14-15 --- 17-21
Y	US 2006/0023706 A1 (VARMA, A. et al.) 02 February 2006; paragraph [0039].	3
Y	US 2007/0055498 A1 (KAPILOW, D. et al.) 08 March 2007; paragraphs [0026], [0090].	4-8
Y	US 5,907,822 A (PRIETO, JR. J.) 25 May 1999; column 7, lines 27-41; column 8, lines 35-39.	8
Y	WO 2006/079349 A1 (SONORIT APS) 03 August 2006; page 12, line 32 - page 13, line 4; page 15, lines 28-31.	9, 13, 21

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 July 2015 (10.07.2015)	24 JUL 2015
Name and mailing address of the ISA/ Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US15/18692

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2002/0133764 A1 (WANG, Y.) 19 September 2002; paragraph [0039].	11-12
Y	US 5,758,319 A (KNITTLE, C.) 26 May 1998; column 4, lines 5-33.	17-20
A	US 6,633,846 B1 (BENNETT, I. et al.) 14 October 2003; entire document.	1-21

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US15/18692

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Group I: Claims 1-9; Group II: Claims 10-13; Group III: Claims 14-21

-***-Please see extra sheet-***-

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US15/18692

-***-Continued from Box No. III - Observations where unity of invention is lacking-***-

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fee must be paid.

Group I: Claims 1-9 are directed toward a method to correct for packet loss in an audio signal in an automatic speech recognition system comprising reconstructing the signal using linear interpolation from preceding and succeeding frames of audio.
Group II: Claims 10-13 are directed toward a method to correct for packet loss in an audio signal in automatic speech recognition systems comprising rescoring of phoneme probability calculations for a section of the audio signal.
Group III: Claims 14-21 are directed toward a method to correct for packet loss in an audio signal in automatic speech recognition systems comprising maximum likelihood prediction.

The inventions listed as Groups I-III do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

The special technical features of Group I include determining spectral vectors; maintaining a sliding buffer; generating the linear interpolation; tag frames within the packet loss event using energy; and inserting values for the packet lock event into the sliding buffer, which are not present in Groups II-III.

The special technical features of Group II include accumulating probabilities for a given word in a lexicon; determining when the probabilities are statistically significant; reporting matches; normalizing values of probabilities that are not statistically significant; and rescoring the probabilities for the given word, which are not present in Groups I and III.

The special technical features of Group III include determining Hidden Markov Model state occupation probabilities in an acoustic model within the system prior to a packet loss event; and predicting feature values of packets in the packet loss section of the audio signal based on the determined Hidden Markov Model state occupation probabilities, which are not present in Groups I-II.

The common technical feature shared by Groups I-III is a method to correct for packet loss in an audio signal in automatic speech recognition systems, comprising: determining probabilities. However, these common features are previously disclosed by US 2009/0112585 A1 to COX et al. (hereinafter "Cox"). Cox discloses a method to correct for packet loss in an audio signal in automatic speech recognition systems (recognizing a stream of speech and including error correction bits in packets; Abstract and paragraph [0009]), comprising: determining probabilities (classify a speech pattern corresponding to a recognition model based on the probability determined for each model; paragraph [0023]).

Since the common technical features are previously disclosed by the Cox reference, these common features are not special and so Groups I-III lack unity.