2007 IEEE/RSJ International Conference on Intelligent Robots and Systems,
San Diego, USA, October 2007

A Unified Approach to Speech Production and Recognition Based on Articulatory Motor Representations

Jonas Hörnstein and José Santos-Victor

***Abstract*— We present a unified approach for speech production and recognition based on articulatory motor representations. The approach is inspired by the Motor theory and the discovery of Mirror neurons, and use motor representations for both reproduction and recognition of speech. A model of the vocal tract is used to create sound and the created sound is then mapped back to the motor representation using a neural network. To learn the map we mimic the behavior of a child that uses a combination of babbling and interaction with its caregiver to learn how to speak. Several different phases of babbling and interaction are identified and described. These help to overcome the inversion problem. The approach has been implemented on a humanoid robot, which has successfully learned to pronounce Swedish and Portuguese vowels. We have also studied how the different phases of babbling and interaction effect the error of the map and the achieved recognition rate when presented with vowels from different subjects. Finally we compare the recognition rates obtained using motor space with recognition rates obtained by directly using the acoustic parameters.**

I. INTRODUCTION

Speech is an important and powerful tool for interaction between humans, and is also a promising method for human-robot communication. To be able to communicate by speech, both partners involved in the communication have to share some common phonemes and have the capability of both vocalizing those as well as recognizing the phonemes as they are vocalized by others. The tasks of finding common phonemes, learning how to vocalize those, and recognition of phonemes are usually handled separately. However, there are reasons to believe that these mechanisms should not be treated independently, and that there can be advantages in handling these by a unified approach. Findings in neuroscience research has shown an increased activity in the tongue muscles when listening to words that requires large tongue movements [1]. This leads to believe that the motor area is involved not only in the task of production, but also in that of recognition. Earlier work including neurophysiological studies of monkeys have shown a similar relationship between visual stimulation and the activation of premotor neurons [2]. Those neurons fire both when executing a motor command and when being presented with an action that involves the same motor command.

This work was partially supported by EU Project CONTACT and by the Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding) through the POS Conhecimento Program that includes FEDER funds.

J. Hörnstein is with Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal jhornstein@isr.ist.utl.pt

J. Santos-Victor is with Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal jasv@isr.ist.utl.pt

Some work has already been done in order to verify the usability of these findings in the area of robotics. More specifically it has been shown that action recognition becomes significantly easier when performed in motor space rather than directly in sensor space [3]. This further motivates the simultaneous study of speech production and recognition. While there are existing work using a mirror neuron inspired approach in speech production, like the DIVA model [4], we use a slightly different approach. They use babbling to learn the auditory-motor maps, but the speech recognition and the motion planning are performed in the auditory space. In this work we directly use motor space to perform the speech recognition as we specifically want to investigate he usability of motor space not only for speech production, but also for speech recognition.

The main contribution of this work is a unified approach that can be used for learning to produce speech, extract useful phonemes, as well as for speech recognition. There are two main components in this approach. One is a flexible architecture where we identify the units involved in speech production and recognition, which make it possible to easy to exchange individual units in order to try different types of methods. The other main component is the different phases of self-exploration and interaction that is used to learn the maps. These learning mechanisms are inspired by the way a child learns to speak through babbling and through the interaction with its caregivers. This makes it possible to solve the inversion problem caused by the many to one relationship between the articulator positions and the produced sound.

The rest of the paper is organized as follows. In section 2 we describe the architecture used for learning speech production and recognition. In section 3 we describe the learning mechanisms used for speech production along with some results and in section 4 we give some results for speech recognition. Conclusions are given in section 5.

II. SPEECH SYSTEM ARCHITECTURE

An overview of the architecture used in this work is shown in Figure 1. The architecture consists of one speech production unit, an auditory sensor unit, a sound-motor map, and a speech recognition unit.

A. Speech production unit

The speech production unit consists of a model of the human vocal tract and a position generator. There has been several attempts to build mechanical models of the vocal tract [5][6]. While these can produce some human like sounds they are still pretty limited and there are no commercially

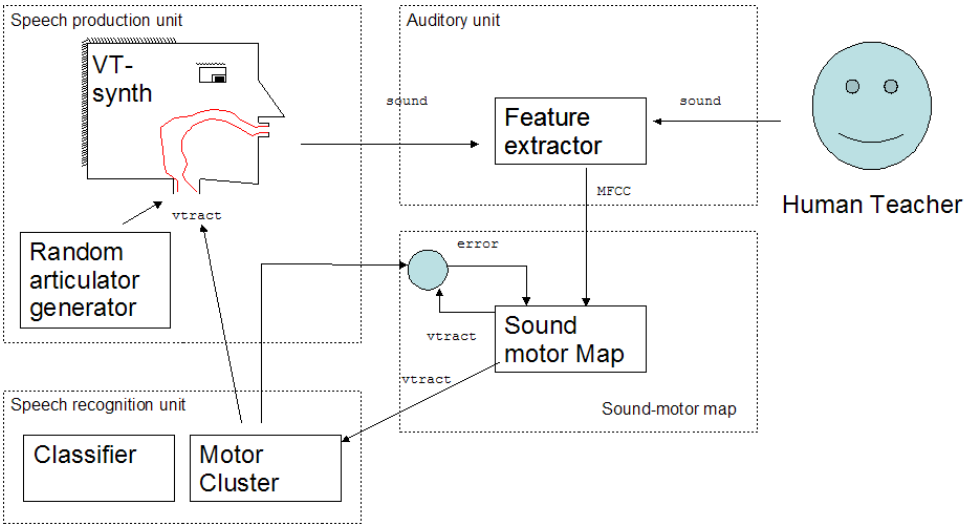

Fig. 1. Speech architecture

available mechanical solutions. The alternative is to simulate the vocal tract with a computer model. Such simulators are typically based on the tube model [7] where the vocal tract is considered to be a number of concatenated tubes with variable diameter. On top of the tube model an articulator is used that calculates the diameter of the tubes for different configurations of the vocalization units. In this work we have chosen to simulate the vocal tract by using vtcals developed by Maeda [8]. This model has been developed by studying x-rays from two women articulating French words, and has six parameters that can be used to control the movements of the vocal tract. One parameter is used for the controlling the position of the yaw, one for the extrusion of the lips, one for lip opening, and three parameters for controlling the position of the tongue. A synthesizer is finally used to calculate the resulting sound from the area function. The positions of the articulators are given by a position generator. This can work either in babbling mode in a direct mode. In the babbling mode it randomly generates position either globally or in the neighborhood of a given position, whereas in the direct mode it receives a position from the speech recognition unit which it simply reproduce.

B. *Auditory unit*

The auditory unit consists of a feature extractor that takes sound as an input and calculates a number of sound features (tonotopic sound representation) that are useful for speech production and recognition. There exists various features that can be used for this. For production and recognition of vowels, formants are commonly used [9]. However, formants only contains information that is useful for vowels so its application is rather narrow. In other related work LPC has been used [10][11]. LPC are more generally applicable than formants, but still requires rather stationary signals to perform good. Even though we are working with vowels in the current study we have decided to use Mel frequency cepstral coefficients (MFCC) [12] as our speech features as these do not put any restrictions on the extendibility of the proposed approach.

C. *Sound-motor map*

The next unit is the sound-motor map. This is responsible for retrieving the vocal tract position from the given sound features. This is a rather well studied inversion problem. This is a non linear problem that becomes extra difficult since several positions of the vocal tract results in the same sound, so it may exist several possible solutions for a given set of features. This is usually solved by introducing some weighting function based on either dynamic restraints or some comfort measurement. In our work is it solved through a combination of babbling and interaction with the caretaker as will be described in the next section. The map itself is usually build by some kind of neural network. In [11] a self organizing neural network is first used to cluster the auditory information before it is passed on to an artificial neural network that map the clustered sound to the vocal tract. We have chosen not to do any clustering in the auditory space, but instead take care of this in the motor space. We have therefore chosen to use a straight forward method where the auditory features are fed directly into an artificial neural network with 20 hidden neurons and trained with back propagation.

D. *Speech recognition unit*

Finally we have the speech recognition unit. This unit stores sounds that it found useful for communication. Useful configurations can either be inserted manually or be learned. Many speech recognition system starts from a given set of phonemes. Other system automatically clusters the information into something that can be considered as pseudo-phonemes [13]. While these techniques are usually applied directly in the auditory space, the same techniques can be used in motor space. In this work we have chosen to extract useful speech units through the interaction with the caregiver

as will be described in the next section. While we only consider vowels in this work, the same approach can be used for other phonemes. The vowels given by the caretaker is stored in a motor cluster. In the current work we only teach the robot one suitable position for each desired sound. The cluster is therefore reduced to a a simple dictionary. The recognition task is handled by the classifier that compares positions given from the sound motor map with the positions stored in the motor cluster. We have implemented two classifiers for the recognition, one that uses Euclidean distance and one that uses Mahalanobis distance to find the nearest neighbor.

The Euclidean distance d_1 is calculated directly as the distance between the given position p and each of articulator positions c_i stored in the cluster.

$$d_1 = \sqrt{(p - c_i)^T (p - c_i)}$$

To calculate the Mahalanobis distance we first calculate the mean value μ_i and the covariance matrix Σ_i for a number of mapped training data for each class c_i in the cluster. The Mahalanobis distance d_2 between a given position p and stored class is then calculated as:

$$d_2 = \sqrt{(p - \mu_i)^T \Sigma_i^{-1} (p - \mu_i)^T}$$

III. Speech production

In this this section we look at how a robot can use the architecture described in the previous section to learn how to vocalize vowels. The method used is inspired by the way children develop their speech through a combination of self-exploration in the form of babbling and through the interaction with a caregiver. Early babbling can be seen as random movements of the articulators, while in the later stages the babbling gets more rhythmic and focused around some stationary points. Here we first do a random babbling and later focus the babbling around the learned vowels.

A. Initial babbling

During the initial babbling random motor positions are generated by the position generator. We generated 10000 random positions vectors for this phase. Each vector contains information about the position of the 6 articulators used in Maedas model. These are then passed on to the speech production unit that calculates the resulting sound. The sound is then fed into the auditory unit that calculates the MFCC and passes these to the sound-motor-map. The sound-motor-map finally tries to map the MFCC back to the original articulator positions that originated the sound, compares it with the correct position given by the random articulator generator, and uses a back-propagation algorithm to update the map. Repeating this will create an initial map between sound and the articulator positions used to create this sound. Figure 2 shows the learning curve for the initial babbling.

Even after a long period of initial babbling a relatively large residual error remains. This is because of the inversion problem described in the introduction. Several articulator positions can result in the same or similar sound. This problem will be resolved in the next steps where the robot in interaction with its caregiver starts to focus on a number of useful articulator positions rather than the whole articulator space.

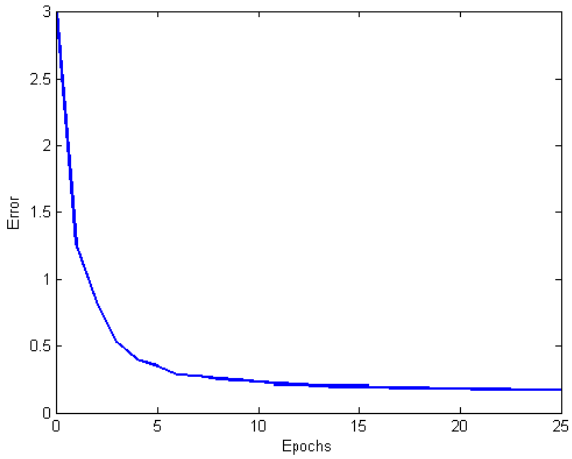

Fig. 2. Average error of the mapped motor representations during initial babbling for each epoch in the training. The error is expressed as the distance between the correct and the mapped motor representation. As a comparison, the distance between two different vowels is typically between 0.2 and 0.5

B. Learning vowels

The second phase can be seen as a parroting behavior where the robot tries to imitate the caregiver using the previously learned map. Since the map at this stage is only trained with the robot's own voice, it will not generalize very well to different voices. This may force the caretaker to change his or her own voice in order to direct the robot. This behavior can also be found in the interaction between a child and its parents, when the parents speak "baby language". There can also be a need to over-articulate, i.e. exaggerate the positions of the articulators in order to overcome flat areas in the maps that are a result of the inversion problem. When two or more articulator positions give the same sound the initial maps tends to be an average of those. However, for vowels the articulator positions are usually naturally biased towards the correct position as the sound is more stable around the correct positions than around the alternative positions. For most of the vowels it was not necessary to adapt the voice too much. Typically between one and ten tentatives were enough to receive a satisfying result. When the caregiver is happy with the sound produced by the robot it gives positive feedback which causes the robot to store the current articulator positions in its cluster. This reinforcement was given though the keyboard in the current implementation, but more sophisticated methods could be used.

Using this technique we have been able to teach the robot complete sets of Swedish and Portuguese vowels. Looking at the articulator positions used by the robot we find that these are similar to those used by a human speaker, see Figure 3.

Fig. 3. Learned articulator positions for Portuguese vowels. The upper left position corresponds to the Portuguese vowel ɔ used in for example the word só. The next position to the right correspondes to the vowel ɛ used in sé, and the following positions corresponds to the vowels: o (sou), a (vá), e (vê), ɨ (pegar), ɐ (pagar), u (mudo), and i (vi).

C. Extended babbling

The robot can now use the learned set of articulator positions as a starting point for further exploration, again using a combination of self-exploration and interaction. In this phase the self-exploration is biased towards the learned articulator positions in order to specialize the maps to these areas, and overcome the problem with ambiguous articulator positions previously described. As a result, during this phase the error in the sound-motor map for the learned vowels is drastically reduced, Figure 4.

A secondary result of this phase is that articulator positions that are not used in what is becoming the robot's mother language, will be less likely to be reached by the map. The more a robot gets accustomed to use one set of articulator positions, the more likely it will be to map the sound to one of those, and the harder it will be for the robot to learn new positions. While this may not be a desired features it is interesting to do a parallell to humans as children that have contact with several languanges during the babbling phases can easily learn to produce sounds from all languanges, while this gets increasingly difficult with age.

D. Gaining speaker invariance

One problem that the robot still has to overcome is the difficulty to map different voices to the correct articulator positions. This again has to be learned through the interaction with one or more caregivers. One possible way to do this is to let the caregiver imitate the robot, using his or her own voice. The robot simply utter one of the sounds it has previously learned and then listen to same utterance produced by the caregiver. It can now update the sound-motor-map using the sound produced by the caregiver and the articulator positions used by the robot. Doing this with several different caregivers will gradually introduce more invariance to speakers.

IV. Speech recognition

Being able to reproduce a sound that closely match the original one doesn't necessarily mean that robot know what it actually said. In this section we deal with the problem to classify a given sound as one of the previously learned classes, i.e. one of the vowels. Especially we want to test two things. First we study the effect on speech recognition of the different phases of babbling and interaction with the caregiver described in the previous section. Second we want to compare the result of using motor parameters for speech recognition with results from directly using MFCC.

A. Experimental setup

We have performed some initial experiments using 14 speakers (seven males and seven females) reading words

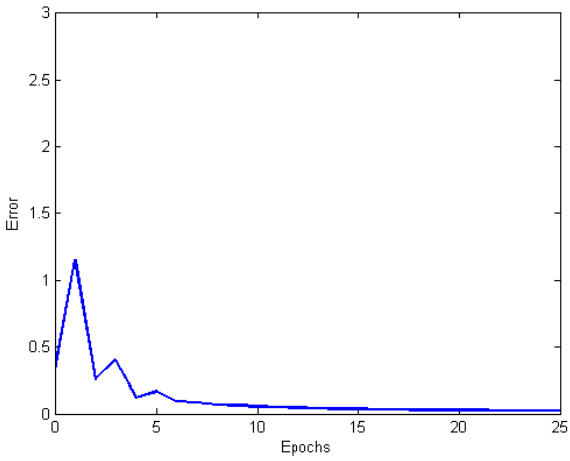

Fig. 4. Average error of the mapped motor representations for the learned vowels as a function of the number of epochs, when using local babbling.

that included the nine Portuguese vowels previously learned by the robot. We used the vowels from seven speakers for training and the other seven for testing. Each speaker read the words several times, and the vowels were hand labled with a number 1 to 9. The amplitude of the sound was normalized and each vowel was then divided into 30 ms windows with 50% overlap. Each window was then treated as individual data which resulted in a training set of 2428 samples, and a test set of 1694 samples.

We performed two different experiments. The first experiment was made to study the effect of babbling. During this experiment we measured the average sum of square error between the mapped vowels and correct articulator position. We also used the Euclidean distance to classify each mapped vowel to the closest of the stored positions and calculated the percentage of correctly classified vowels.

In the second experiment we wanted to test how well the mapped motor representations perform in comparison with using the MFCC directly. For this experiment we used the two classifiers described in section 2, which calculate the nearest class using Euclidean and Mahalanobis distances respectively. Two instances of each classifier were created. One of the instances used MFCC and the other used motor representations. The Euclidean distance was calculated directly using the stored vowels. For the Mahalanobis distance we first calculated the mean and covariance matrix for the training vowels. In the second experiment we also added noise to the test vowels in order to measure the sensitiveness to noise of the two classifiers.

B. Effect of babbling

In the first experiment we studied the effect of the various phases of babbling described in the previous section. During the first phase we only performed initial babbling where the robot randomly moved its articulators and listen to the produced sound. During this training phase the robot had not heard any human sound and as expected the recognition ratio at this stage is very low, table 1. Notice that, due to the inversion problem, the recognition rate for the robot vowels is also very low at this stage.

TABLE I

RESULTS AFTER INITIAL BABBLING

Results	Sum of square distance	recognition rate
Robot vowels	3,4699	28,18%
Human vowels	9,7475	17,53%

TABLE II

RESULTS AFTER EXTENDED BABBLING

Results	Sum of square distance	recognition rate
Robot vowels	0,3195	84,44%
Human vowels	1,9839	22,22%

During the second phase the robot has retrieved a set of vowels and use these for a local babbling. While the robot at this stage has already heard at least the voice of one caregiver, we only used the robot sound from the local babbling for training. The results are shown in table 2. Here it can be seen that the recognition rate of robot vowels has drastically increased. In fact, by restricting the local babbling to a very narrow neighborhood around the vowels we could easily obtain 100% recognition rate for the robot vowels. However, that would make the map too specialized and it would not generalize will to human vowels. During this phase we therefore generated motor positions using random distributions with mean valus equal to the learned vowels and a standard deviation of 0,1.

Finally we use the set with human training vowels to train the network. The results of this is shown in table 3. As expected the recognition ratio is increased for the human vowels. As we only use human vowels for training in this step, the map will get increasingly more specialized on human vowels and the recognition rate for the robot vowels is therefore decreasing during this stage.

C. Comparison between recognition ratios for audio and motor space

A second experiment was performed where we compared the performance of the speech recognition calculated using the motor representation with the speech recognition ratio calculated using MFCC. We also tested the sensitiveness to noise for these. Two different measurements were used for the classification, Euclidean distance and Mahalanobis distance. The results from these classifications are shown in table 4 and table 5 respectively.

TABLE III

RESULTS AFTER INTERACTION

Results	Sum of square distance	recognition rate
Robot vowels	0,5568	49,10%
Human vowels	0,5150	57,67%

TABLE IV

RECOGNITION RATES USING EUCLIDEAN DISTANCE

Results	Motor positions	MFCC
no noise	57,67%	20,60%
noise 15%	51,48%	24,56%
noise 30%	41,15%	24,09%

TABLE V

RECOGNITION RATES USING MAHALANOBIS DISTANCE

Results	Motor positions	MFCC
no noise	59,27%	56,85%
noise 15%	45,99%	48,64%
noise 30%	35,95%	37,66%

As seen the Euclidean distance gives very poor results when used with the MFCC. This is because we try to directly map the sound of the humans to the sound produced by the robot. However, the acoustic space of the robot is quite different from the acoustic space of humans so even if the sound is perceived as being the same vowel by a human listener, by looking at the direct values of the MFCC in most cases we would conclude that they are not. This difference in acoustic space is not something that just robots have to worry about. Children have to deal with exactly the same problem when learning to speak as their acoustic space is very far from that of their adult caregivers. By mapping the MFCC to motor representation we can simplify the classification and greatly improve the recognition rate when using a simple Euclidean distance.

On the other hand, if we use a more advanced classifier as the one based on the Mahalanobis distance, which uses the human training vowels to recognise the test vowels rather than the sound produced by the robot, the recognition rates for MFCC get close to those optained by using motor representation.

V. CONCLUSIONS

We have presented an unified approach for speech production and recognition. This uses a combination of babbling and interaction with a caregiver to automatically learn both how to pronounce vowels and to recognize vowels pronounced by other subjects.

It is found that using self-exploration, i.e. babbling, alone is not enough to create a good map between acoustic and articulatory space. This is because several articulatory positions result in the same articulated sound. However, we find that the map is good enough to let the caregiver guide the robot towards the correct articulator positions by changing his or her voice. We have successfully taught the robot both Swedish and Portuguese vowels.

Once the robot has learned a number of useful positions it can concentrate the babbling to the neighborhood of those positions. This way the robot can overcome the inversion problem and correctly learn to map its own sound to its articulator positions. However, there is still a big difference between the sound produced by the robot for a specific vowel, and sound produced by humans. Also, sound produced by different individuals, or by the same individual under different conditions, also result in a variety of different sounds for the same vowel. To gain some speaker invariance the robot has to interact with several different speakers. We let the robot interact with seven speakers and then used the speech of seven different individuals to test the recognition rate. It was found that the recognition rate based on the motor representations were as good or better than the rates obtained by directly using MFCC. Also, in the case of the motor representations, it was possible to use a very simple classifier.

While it is early to draw any general conclusions about the performance of using motor representations compared to directly using acoustical features, these results are encouraging and future work will include validating the results on other data sets, and also extending the data sets to include other phonemes and words.

REFERENCES

[1] Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G., "Speech listening specifically modulates the excitability of tongue muscles: a TMS study", European Journal of Neuroscience, Vol 15, pp. 399-402, 2002

[2] Gallese, V. and Fadiga, L. and Fogassi, L. and Rizzolatti, G. "Action Recognition in the Premotor Cortex", Brain, 199:593-609, 1996

[3] Lopes, M., Santos-Victor, J. "Visual learning by imitation with motor representations" *IEEE - Transactions on Systems, Man, and Cybernetics - Part b: Cybernetics*, 35(3), 2005

[4] Guenther, F. H., Ghosh, S. S., and Tourville, J. A., "Neural modeling and imaging of the cortical interactions underlying syllable production", Brain and Language, 96 (3), pp. 280-301

[5] Higashimoto, T. and Sawanda, H., "Speech Production by a Mechanical Model: Construction of a Vocal Tract and Its Control by Neural Network" in proc. International Conference on Robotics and Automation, Washington DC, pp 3858-3863, May 2002

[6] Fukui, K., Nishikawa, K., Kuwae, T., Takanobu, H., Mochida, T., Honda, M., and Takanishi, A., "'Development of a New Human-like Talking Robot for Human Vocal Mimicry", in proc. International Conference on Robotics and Automation, Barcelona, Spain, pp 1437-1442, April 2005

[7] Liljencrants, J. and Fant, G., "Computer program for VT-resonance frequency calculations", STL-QPSR, pp. 15-20, 1975

[8] Maeda, S., "Compensatory articulation during speech: evidence from the analysis and synthesis of vocat-tract shapes using an articulatory model", in Speech production and speech modelling (W. J. Hardcastle and A. Marchal, esd.), pp. 131-149. Boston: Kluwer Academic Publishers

[9] Yoshikawa, Y., Koga, J., Asada, M., Hosoda, K., "Primary Vowel Imitation between Agents with Different Articulation Parameters by Parrot-like Teaching", in proc. Int. Conference on Intelligent Robots and Systems, Las Vegas, Nevada, pp. 149-154, October 2003

[10] Krstulovic, S., "LPC modeling with speech production constraints", in proc. 5th speech production seminar, 2000

[11] Nakamura, M. and Sawada, H., "Talking Robot and the Analysis of Autonomous Voice Acquisition"' in proc. International Conference on Intelligent Robots and Systems, Beijing, China, pp 4684-4689, October 2006

[12] Davis, S. B., Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, speech, and signal processing, Vol. ASSP-28, no. 4, August 1980

[13] Salvi, G., "Ecological Language Acquisition via Incremental Model-Based Clustering", Interspeech 2005, 9th European Conference on Speech Communication and Technology